

UNIFIED APPROACH FOR MINIMIZING COMPOSITE NORMS

N. S. AYBAT* AND G. IYENGAR†

Abstract. We propose a first-order augmented Lagrangian algorithm (FALC) to solve the composite norm minimization problem

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta, \\ \text{subject to} \quad & \|\mathcal{A}(X) - b\|_\gamma \leq \rho, \end{aligned}$$

where $\sigma(X)$ denote the vector of singular values of the matrix $X \in \mathbb{R}^{m \times n}$, the matrix norm $\|\sigma(X)\|_\alpha$ denotes either the Frobenius, the nuclear, or the ℓ_2 -operator norm of the matrix X , the vector norms $\|\cdot\|_\beta, \|\cdot\|_\gamma$ denote either the ℓ_1 -norm, ℓ_2 -norm or the ℓ_∞ -norm; and $\mathcal{A}(\cdot), \mathcal{C}(\cdot)$ are linear operators from $\mathbb{R}^{m \times n}$ to vector spaces of appropriate dimensions. This formulation includes as special cases problems such as basis pursuit, matrix completion, robust PCA, and stable PCA. Thus, the FALC is able to solve all these problems in a unified manner.

FALC solves this semidefinite optimization problem by inexactly solving a sequence of problems of the form

$$\min \left\{ \begin{aligned} & \lambda^{(k)} \mu_1 \|X\|_\alpha + \frac{1}{2} \|\mathcal{A}(X) + y - b - \lambda^{(k)} \theta_1^{(k)}\|_2^2 \\ & + \lambda^{(k)} \mu_2 \|s\|_\beta + \frac{1}{2} \|\mathcal{C}(X) + s - d - \lambda^{(k)} \theta_2^{(k)}\|_2^2 \end{aligned} : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho \right\},$$

for an appropriately chosen sequence of multipliers $\{\lambda^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}\}_{k \in \mathbb{Z}_+}$. Each of these non-smooth subproblems are solved inexactly using Algorithm 3 in [33] where each update involves computing at most one singular value decomposition (SVD). We show that FALC converges to the optimal solution X_* of the composite norm minimization problem whenever the optimal solution is unique. We also show that there exists a priori fixed sequence $\{\lambda^{(k)}\}_{k \in \mathbb{Z}_+}$ such that for all $\epsilon > 0$, iterates $X^{(k)}$ computed by FALC are ϵ -feasible and ϵ -optimal after $\mathcal{O}(\log(\frac{1}{\epsilon}))$ iterations, which requires $\mathcal{O}(\frac{1}{\epsilon})$ operations in total where the complexity of each operation is dominated by computing a singular value decomposition. We also show that FALC can be extended to solve problems where, in addition to the constraints above, we have constraints of the form $\mathcal{F}(X) \preceq G$ where $\mathcal{F}(\cdot)$ is linear operator and \preceq denotes the partial order with respect to the cone of positive semidefinite matrices. All the convergence properties of FALC continue to hold for this more general problem.

1. Introduction. We propose a first-order augmented Lagrangian algorithm (FALC) to solve the class of composite norm minimization problems defined as follows:

$$\min_{X \in \mathbb{R}^{m \times n}} \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta \text{ subject to } \|\mathcal{A}(X) - b\|_\gamma \leq \rho, \quad (1.1)$$

where $\sigma(X) \in \mathbb{R}_+^{\min\{m,n\}}$ denotes the vector of singular values of the matrix $X \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^q$, $d \in \mathbb{R}^p$, $\mathcal{A} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^q$, $q < mn$, and $\mathcal{C} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^p$, $p < mn$, are linear operators and $\alpha, \beta, \gamma \in \{1, 2, \infty\}$. For $\alpha \in \{1, 2, \infty\}$ the vector norm $\|\cdot\|_\alpha$ denotes the ℓ_1 -norm, ℓ_2 -norm or the ℓ_∞ -norm, respectively. Since the Nuclear norm $\|X\|_* = \|\sigma(X)\|_1$, the Frobenius norm $\|X\|_F = \|\sigma(X)\|_2$, and the ℓ_2 -operator norm $\|X\|_2 = \|\sigma(X)\|_\infty$, the term $\|\sigma(X)\|_\alpha$ denotes either the nuclear, the Frobenius, or the ℓ_2 -operator norm. We assume that \mathcal{A} has full rank – we do not need this constraint on the operator \mathcal{C} . Although we focus on establishing the properties of FALC for problems of the form (1.1), we show in Section 5 that our proposed framework extends to a much larger class of problems. We show below that many well studied optimization problems are special cases of (1.1).

Nuclear norm-minimization. The special case with $\alpha = 1$, i.e. $\|\sigma(X)\|_1 = \|X\|_*$, $\mu_2 = 0$, and $\rho = 0$, is known as *nuclear norm minimization problem*

$$\min_{X \in \mathbb{R}^{m \times n}} \|X\|_* \text{ subject to } \mathcal{A}(X) = b. \quad (1.2)$$

Nuclear norm minimization problem is a convex approximation for the NP-hard rank minimization problem $\min_{X \in \mathbb{R}^{m \times n}} \{\text{rank}(X) : \mathcal{A}(X) = b\}$, where $\text{rank}(X)$ denotes the rank of $X \in \mathbb{R}^{m \times n}$. Rank minimization arises in many different contexts, e.g. system identification [28], optimal control [15, 16, 14], low-dimensional embedding in Euclidean space [27], and matrix completion. Matrix completion is the special case where the operator \mathcal{A} picks a subset of the matrix elements, i.e the linear constraints are of the form $X_{ij} = M_{ij}$ for $(i, j) \in \Omega$. The Netflix prize problem [30] is an example of the matrix completion problem. Recently, Recht

*IEOR Department, Columbia University. Email: nsa2106@columbia.edu

†IEOR Department, Columbia University. Email: gi10@columbia.edu

et al. [31] have shown that when the linear operator $\mathcal{A} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^q$ satisfies some regularity properties, and the number of measurements $q = \mathcal{O}(r(m+n) \ln(mn))$, the optimal solution of the SDP (1.2) is the optimal solution of the rank minimization problem with very high probability. Thus, FALC can be used to approximately solve rank minimization problems. For existing algorithmic methodologies for solving the nuclear norm minimization problem see [4, 18, 25, 26, 29, 32] and references therein.

Basis-pursuit problem. The special case of (1.1) with $\beta = 1$, $\mu_1 = 0$, $\rho = 0$, $C(X) = X \in \mathbb{R}^{n \times 1}$ and $d = 0$, is known as the *basis pursuit problem*

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \text{ subject to } Ax = b, \quad (1.3)$$

where $A \in \mathbb{R}^{q \times n}$ and $b \in \mathbb{R}^q$. LPs of the form (1.3) have recently attracted a lot of attention since they appear in the context of a new signal processing paradigm known as *compressive sensing* (CS) [5, 6, 7, 13]. The goal in CS is to recover a sparse signal x_0 from a small set of linear measurements or transform values $b = Ax_0$, or equivalently, to solve the NP-hard ℓ_0 -minimization problem

$$\min_{x \in \mathbb{R}^n} \|x\|_0 \text{ subject to } Ax = b, \quad (1.4)$$

where the ℓ_0 -norm $\|x\|_0 = \sum_{i=1}^n \mathbf{1}(x_i \neq 0)$. Recently, Candés, Romberg and Tao [5, 6, 7] and Donoho [13] have shown that, when the target signal x_0 is s -sparse, i.e. only s of the n components are non-zeros, the matrix $A \in \mathbb{R}^{q \times n}$ has $q = \mathcal{O}(s \ln(n))$ and is chosen randomly according to a specified set of distributions, the sparse target signal x_0 is the optimal solution of the LP (1.3) with very high probability. Thus, x_0 can be recovered by solving an LP, and therefore, in theory signal recovery is very efficient. In practice, however, solving such LPs is hard because the matrix A in (1.3) is large and dense, and in addition these LPs are often ill-conditioned. Thus, general purpose simplex-based LP solvers are not able to efficiently solve (1.3). The measurement matrix A in CS applications has a lot of structure, in particular the matrix-vector multiplication Ax and $A^T y$ can be computed efficiently. Recently, a number of different algorithms have been proposed to exploit this structural fact to efficiently solve (1.3) [1, 2, 12, 17, 19, 20, 22, 34, 35, 36].

Principal component pursuit. The special case of (1.1) with $\alpha = 1$, i.e. $\|\sigma(X)\|_\alpha = \|X\|_*$ denoting the nuclear norm, $\beta = 1$, $\mu_1 = 1$, $\mu_2 > 0$, $\mathcal{A} = \mathbf{0}$, $b = 0$, $\rho = 0$ and the operator $\mathcal{C} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$ such that $\mathcal{C}(X) = \mathbf{vec}(X)$, where $\mathbf{vec}(X)$ is vector obtained by stacking the columns of $X \in \mathbb{R}^{m \times n}$ in order, is the *principal component pursuit problem*

$$\min_{X \in \mathbb{R}^{m \times n}} \|X\|_* + \mu_2 \|\mathbf{vec}(X) - d\|_1. \quad (1.5)$$

In [8, 25], it is shown that when the data matrix $D \in \mathbb{R}^{m \times n}$ is of the form $D = X_0 + S_0$, where X_0 is a low rank matrix and S_0 is a sparse matrix, then one can recover the low rank and sparse components of D by solving the problem given in (1.5) for an appropriately chosen μ_2 . In [37], it was shown that the recovery is still possible even when the data matrix is corrupted with a dense error matrix. When the data matrix D is of the form $D = X_0 + S_0 + Y_0$, where X_0 is a low rank matrix, S_0 is a sparse matrix and $\{(Y_0)_{ij}\}$ is independent and identically distributed for all i, j such that $\|Y_0\|_F \leq \rho$, solving the *stable principal component pursuit problem*

$$\begin{aligned} \min_{X, S \in \mathbb{R}^{m \times n}} \quad & \|X\|_* + \mu_2 \|\mathbf{vec}(S)\|_1, \\ \text{subject to} \quad & \|X + S - D\|_F \leq \rho. \end{aligned} \quad (1.6)$$

produces (X_*, S_*) such that $\|X_* - X_0\|_F^2 + \|S_* - S_0\|_F^2 \leq Cmn\rho^2$ for some constant C with high probability. Principal component pursuit and stable principle component pursuit both have applications in video surveillance and face recognition. For existing algorithmic approaches to solving principal component pursuit see [8, 18, 25, 26, 37] and references therein.

Matrix completion problems with semidefinite constraints. In Section 5 we show that FALC can be extended to solve can solve a larger class of optimization problems of the form

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta + \langle R, X \rangle, \\ \text{subject to} \quad & \|\mathcal{A}(X) - b\|_\gamma \leq \rho, \\ & \mathcal{G}(X) = h, \\ & \mathcal{F}(X) \preceq G, \end{aligned} \quad (1.7)$$

where $X_0, R \in \mathbb{R}^{m \times n}$, \mathcal{G} is a linear operator that maps X to a vector, \mathcal{F} is linear operator that maps X to a symmetric matrix, and \preceq denotes the partial order induced by the cone of positive semidefinite matrices.

In the *sparse* PCA problem

$$\min_{x \in \mathbb{R}^n} \quad \|\Sigma - xx^T\|_F \text{ subject to } \|x\|_0 \leq s, \quad (1.8)$$

the goal is to compute an s -sparse vector x that is “close” to the eigenvector corresponding to the largest eigenvalue of the positive semidefinite matrix Σ . The optimization problem (1.8) is not convex, and is, therefore, hard to solve. Let $X = xx^T$. Then (1.8) is equivalent to $\min_{X \in \mathbb{R}^{m \times n}} \{\|X - \Sigma\|_F : \|\mathbf{vec}(X)\|_0 \leq s^2, \text{rank}(X) = 1, X \succeq 0\}$. Since $\|X\|_*$ is the tightest convex upper bound for $\text{rank}(X)$, and $\|X\|_* = \text{Tr}(X)$ for positive semidefinite matrices, the relaxed problem

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times n}} \quad & \|X - \Sigma\|_F + \mu \|\mathbf{vec}(X)\|_1 + \nu \langle I, X \rangle, \\ \text{subject to} \quad & X \succeq 0. \end{aligned} \quad (1.9)$$

is a convex approximation for (1.8), where μ and ν control the sparsity on the entries and the singular values of X , respectively. See [9, 10, 21] for existing approaches for solving the sparse PCA problem.

Problems of the form (1.7) also appear in signal shaping applications. One such problem is the design of the optimal acquisition basis for radar applications. For simplicity assume that we are in a 1-D setting, and we discretize the space. Let $x \in \mathbb{R}^L$ denote the unknown locations of objects (i.e. reflectors). Let $d(t)$ denote the signal transmitted by the radar. Then the received signal $y \in \mathbb{R}^N$ is of the form

$$y(t) = \sum_k d(t - \tau_k) x_k + \eta_t, \quad t = 1, \dots, N.$$

where τ_k denotes the round-trip delay corresponding to the k -th discrete location, and $\eta_t \sim \mathcal{N}(0, \sigma^2)$ denotes the receiver noise. Thus, $y = Dx + \eta \in \mathbb{R}^N$, where the columns of $D \in \mathbb{R}^{N \times L}$ are impulse response of the channel for different delays, and $\eta \sim \mathcal{N}(0, \sigma^2 I)$. Typically, x is very sparse, i.e. $\|x\|_0 \ll L$, therefore, one could recover x by solving the CS-like LP $\min\{\|x\|_1 : \|y - Dx\|_2 \leq \sigma\}$. The power cost of this approach is likely to be high since the matched filter and the analog-to-digital converter (ADC) has to run at a very high rate to generate $y(t)$ for all t . Since x is sufficiently sparse, it is possible that x can be recovered from a lower dimensional projection Wy of the observations y by solving the LP $\min\{\|x\|_1 : W Dx = Wy\}$ for some $W \in \mathbb{R}^{M \times N}$, where $M < N$. Such a projection saves device power because the matrix multiplication Wy is implemented as a matched filter in the analog domain and the ADC only needs to convert the product. For good performance of this strategy, one requires the following properties.

- (i) small row dimension of W : this ensures a low dimensional transformed signal Wy .
 - (ii) small *mutual incoherence* $\max\{|(D^T W^T W D)_{ij}| : \text{diag}(D^T W^T W D) = I\}$ of the measurement matrix WD : this ensures that x can be reliably recovered by the LP.
 - (iii) small noise power $\sigma^2 \text{Tr}(W^T W)$ of the compressed signal Wy .
- Let $K = W^T W \succeq 0$. Since the nuclear $\|K\|_* = \text{Tr}(K)$ is good approximation for $\text{rank}(K) = \text{rank}(W)$, a good projection matrix W can be computed by solving the SDP

$$\begin{aligned} \min_{K \in \mathbb{R}^{N \times N}} \quad & \mu_1 \langle I, K \rangle + \mu_2 \|\mathbf{vec}_{off}(D^T K D)\|_\infty, \\ \text{subject to} \quad & \text{diag}(D^T K D) = I \\ & K \succeq 0, \end{aligned} \quad (1.10)$$

where $\text{diag}(X)$ denote a diagonal matrix with entries given by the diagonal elements of X , $\mathbf{vec}_{off}(X) = \mathbf{vec}(X - \text{diag}(X))$, and I is an identity matrix of size N .

1.1. FALC approach and summary of results. The composite norm minimization problem (1.1) can be reformulated as a semidefinite program (SDP), and can, therefore, in theory, be solved efficiently [3]. However, for practical instances the resulting SDPs are large and typically dense. Therefore, interior point based SDP solvers perform very poorly on these instances. Recently a number of different first-order or restricted memory methods have been proposed for solving special cases of (1.1). In the previous section we provide references to the existing literature on algorithmic approaches for solving many of these special cases.

We propose a first-order augmented Lagrangian algorithm (FALC) to solve (1.1) and show that FALC can be extended easily to solve Problem (1.7) with the same complexity guarantees. In Section 2, we establish the convergence properties of FALC for the optimization problem (1.1) and later in Section 5, we briefly describe the extension to the more general problem (1.7).

To obtain separable and efficiently solvable subproblems, we introduce a slack variables s and y , and reformulate (1.1) as

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta, \\ \text{subject to} \quad & \mathcal{C}(X) + s = d, \\ & \mathcal{A}(X) + y = b, \\ & \|y\|_\gamma \leq \rho. \end{aligned} \quad (1.11)$$

We solve (1.11) by inexactly solving a sequence of optimization problems of the form

$$\min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y: \|y\|_\gamma \leq \rho} \left\{ \begin{aligned} & \lambda^{(k)} (\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) \\ & - \lambda^{(k)} (\theta_1^{(k)})^T (\mathcal{A}(X) + y - b) + \frac{1}{2} \|\mathcal{A}(X) + y - b\|_2^2 \\ & - \lambda^{(k)} (\theta_2^{(k)})^T (\mathcal{C}(X) + s - d) + \frac{1}{2} \|\mathcal{C}(X) + s - d\|_2^2 \end{aligned} \right\}, \quad (1.12)$$

for an appropriately chosen sequence of parameters $\{(\lambda^{(k)}, \theta_1^{(k)}, \theta_2^{(k)})\}_{k \in \mathbb{Z}_+}$. We solve these subproblems using Algorithm 3 in [33] where in each update step we need to solve one problem of the form

$$\min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y: \|y\|_\gamma \leq \rho} \left\{ \begin{aligned} & \lambda^{(k)} (\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + \begin{bmatrix} \nabla_X f^{(k)}(\hat{X}, \hat{s}, \hat{y}) \\ \nabla_s f^{(k)}(\hat{X}, \hat{s}, \hat{y}) \\ \nabla_y f^{(k)}(\hat{X}, \hat{s}, \hat{y}) \end{bmatrix}^T \begin{bmatrix} X - \hat{X} \\ s - \hat{s} \\ y - \hat{y} \end{bmatrix} \\ & + \frac{L}{2} \|X - \hat{X}\|_F^2 + \frac{L}{2} \|s - \hat{s}\|_2^2 + \frac{L}{2} \|y - \hat{y}\|_2^2 \end{aligned} \right\}, \quad (1.13)$$

where

$$\begin{aligned} f^{(k)}(X, s, y) = & - \lambda^{(k)} (\theta_1^{(k)})^T (\mathcal{A}(X) + y - b) + \frac{1}{2} \|\mathcal{A}(X) + y - b\|_2^2 \\ & - \lambda^{(k)} (\theta_2^{(k)})^T (\mathcal{C}(X) + s - d) + \frac{1}{2} \|\mathcal{C}(X) + s - d\|_2^2 \end{aligned}$$

denotes the “smooth” part of the objective function in (1.12). Note that (1.13) is *separable* in X , s and y , and it reduces to one vector “shrinkage” [11] (or constrained “shrinkage”, see (2.30)) in s and in y , and one “matrix shrinkage” [29] (or constrained “matrix shrinkage”, see (2.29)) in X .

In this paper we establish the following properties for the FALC algorithm.

(a) Every limit point \bar{X} of the FALC iterates $\{X^{(k)}\}$ is an optimal solution of (1.1), i.e.

$$\bar{X} \in \operatorname{argmin}_{X \in \mathbb{R}^{m \times n}} \left\{ \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta : \|\mathcal{A}(X) - b\|_\gamma \leq \rho \right\}.$$

(b) Suppose (1.1) has a unique optimal solution X_* . There exist a priori fixed sequence $\{\lambda^{(k)} : k \geq 1\}$ such that for *all* $\epsilon > 0$, iterates $X^{(k)}$ computed by FALC are ϵ -feasible and ϵ -optimal, i.e.

$$\begin{aligned} & \|\mathcal{A}(X^{(k)}) + y^{(k)} - b\|_2 \leq \epsilon, \quad \|y^{(k)}\|_\gamma \leq \rho \\ & \left| \left(\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta \right) - \left(\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta \right) \right| \leq \epsilon, \end{aligned}$$

after $\mathcal{O}(\epsilon^{-1})$ iterations, where the complexity of each iteration is $\mathcal{O}(\min\{nm^2, n^2m\})$.

This paper is organized as follows. In Section 2 we prove the main convergence results for FALC and in Section 3 we discuss all the implementation details of FALC. In Section 4 we report the results from our numerical experiments comparing FALC with other algorithms to solve principle component pursuit problems. Finally, in Section 5, we briefly discuss the general problem (1.7) and conclude.

OUTLINE OF FIRST-ORDER AUGMENTED LAGRANGIAN ALGORITHM

input: multipliers $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k \in \mathbb{Z}_+}$, $X^{(0)} \in \mathbb{R}^{m \times n}$, $s^{(0)} \in \mathbb{R}^p$, $y^{(0)} \in \mathbb{R}^q$
 $\eta = \mu_1 \|\sigma(X^{(0)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(0)}) - d\|_\beta$
 $\theta_1^{(1)} = 0$, $\theta_2^{(1)} = 0$, $k \leftarrow 0$
while (Stopping Criterion not true)
 do
 $k \leftarrow k + 1$
 $f^{(k)}(X, s, y) := \frac{1}{2} \|\mathcal{A}(X) + y - b - \lambda^{(k)} \theta_1^{(k)}\|_2^2 + \frac{1}{2} \|\mathcal{C}(X) + s - d - \lambda^{(k)} \theta_2^{(k)}\|_2^2$
 $P^{(k)}(X, s, y) := \lambda^{(k)} (\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f^{(k)}(X, s, y)$
 $h^{(k)}(X, s, y) := \frac{1}{2} \|X - X^{(k-1)}\|_F^2 + \frac{1}{2} \|s - s^{(k-1)}\|_2^2 + \frac{1}{2} \|y - y^{(k-1)}\|_2^2$
 $\eta^{(k)} \leftarrow \eta + \frac{\lambda^{(k)}}{2} (\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2)$
1 Use Algorithm 3 in [33] with $h^{(k)}(X, s, y)$ and initial iterate $(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})$
 compute $(X^{(k)}, s^{(k)}, y^{(k)})$ **such that**
 either
 $P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) \leq \inf\{P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\} + \epsilon^{(k)}$
 or
 $\sqrt{\|G\|_F^2 + \|g\|_2^2} \leq \tau^{(k)}$, for some $(G, g) \in \partial_{X,s} P^{(k)}(\cdot, \cdot, \cdot)|_{(X^{(k)}, s^{(k)}, y^{(k)})}$ **and**
 $\rho \|\nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_{\gamma^*} + \nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T y^{(k)} \leq \xi^{(k)}$,
 with $\mu_1 \|\sigma(X^{(k)})\|_\alpha \leq \eta^{(k)}$
2 $\theta_1^{(k+1)} \leftarrow \theta_1^{(k)} - \frac{\mathcal{A}(X^{(k)}) + y^{(k)} - b}{\lambda^{(k)}}$
3 $\theta_2^{(k+1)} \leftarrow \theta_2^{(k)} - \frac{\mathcal{C}(X^{(k)}) + s^{(k)} - d}{\lambda^{(k)}}$
return $(X^{(k)}, s^{(k)}, y^{(k)})$

FIG. 2.1. Outline of First-Order Augmented Lagrangian Algorithm (FALC)

2. First-order Augmented Lagrangian Algorithm for Composite Norm Minimization. The linear maps \mathcal{A} and \mathcal{C} in (1.1) can be represented as $\mathcal{A}(X) = A \text{vec}(X)$ and $\mathcal{C}(X) = C \text{vec}(X)$, where $A \in \mathbb{R}^{q \times mn}$ and $C \in \mathbb{R}^{p \times mn}$. By completing squares, it is easy to see that (1.12) is equivalent to

$$\min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y: \|y\|_\gamma \leq \rho} \left\{ P^{(k)}(X, s, y) \right\}, \quad (2.1)$$

where

$$\begin{aligned} P^{(k)}(X, s, y) &= \lambda^{(k)} (\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f^{(k)}(X, s, y), \\ f^{(k)}(X, s, y) &= \frac{1}{2} \|\mathcal{A}(X) + y - b - \lambda^{(k)} \theta_1^{(k)}\|_2^2 + \frac{1}{2} \|\mathcal{C}(X) + s - d - \lambda^{(k)} \theta_2^{(k)}\|_2^2. \end{aligned} \quad (2.2)$$

We denote the optimal solution of (1.1) by X_* and, for all $k \geq 1$, we denote the optimal solution of the augmented Lagrangian sub-problem (2.1) by $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})$, i.e. $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) = \text{argmin}\{P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$.

The outline of Algorithm FALC is displayed in Figure 2.1. In Algorithm FALC we use Algorithm 3 in [33] with the prox function $h^{(k)}(X, s) = \frac{1}{2} \|X - X^{(k-1)}\|_F^2 + \frac{1}{2} \|s - s^{(k-1)}\|_2^2 + \frac{1}{2} \|y - y^{(k-1)}\|_2^2$ and initial iterate $(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})$ to compute a new iterate $(X^{(k)}, s^{(k)}, y^{(k)})$ such that one of the following two stopping conditions hold:

$$\begin{aligned} (a) \quad & P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) \leq \min\{P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y: \|y\|_\gamma \leq \rho\} + \epsilon^{(k)} \\ (b) \quad & \sqrt{\|G\|_F^2 + \|g\|_2^2} \leq \tau^{(k)}, \text{ for some } (G, g) \in \partial_{X,s} P^{(k)}(\cdot, \cdot, \cdot)|_{(X^{(k)}, s^{(k)}, y^{(k)})}, \mu_1 \|\sigma(X^{(k)})\|_\alpha \leq \eta^{(k)} \\ & \text{and } \rho \|\nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_{\gamma^*} + \nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T y^{(k)} \leq \xi^{(k)} \end{aligned} \quad (2.3)$$

where $\|\cdot\|_{\gamma^*}$ denotes the dual norm of $\|\cdot\|_\gamma$, $\partial_{X,s} P^{(k)}(\cdot, \cdot, \cdot)|_{(X^{(k)}, s^{(k)}, y^{(k)})}$ denotes the set of partial subgradients of the function $P^{(k)}$ at $(X^{(k)}, s^{(k)}, y^{(k)})$, $\eta := \mu_1 \|\sigma(X^{(0)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(0)}) - d\|_\beta$ for some $X^{(0)} \in \mathbb{R}^{m \times n}$ such that $\mathcal{A}(X^{(0)}) = b$ and $\eta^{(k)} := \eta + \frac{\lambda^{(k)}}{2} (\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2)$.

Since $f^{(k)}(X, s, y)$ is a proper, lower-semicontinuous (lsc), convex function with a Lipschitz continuous gradient, $\nabla f^{(k)}$, the results in [33] guarantee that we can compute $(X^{(k)}, s^{(k)}, y^{(k)})$ in $\mathcal{O}(\frac{1}{\sqrt{\epsilon^{(k)}}})$ operations. In Lemma 2.3 we prove a uniform bound on the number of operations needed to solve any sub-problem encountered in FALC. In the result below we establish that every limit point of the FALC iterate sequence $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k \in \mathbb{Z}_+}$, is an optimal solution for the composite norm minimization problem. In order to compute bounds on the iterates, we need to introduce *dual* norms. The dual $\|\sigma(\cdot)\|_{\alpha^*}$ of the matrix norm $\|\sigma(\cdot)\|_{\alpha}$ is defined as

$$\|\sigma(X)\|_{\alpha^*} = \max\{\langle W, X \rangle : \|\sigma(W)\|_{\alpha} \leq 1\}.$$

It is easy to establish that α^* is the Hölder conjugate of α , i.e. $\frac{1}{\alpha^*} + \frac{1}{\alpha} = 1$ (see Proposition 2.1 in [31] for details). The dual norm of the vector norm $\|x\|_{\beta}$ is clearly $\|x\|_{\beta^*}$, where β^* is the Hölder conjugate of β , i.e. $\frac{1}{\beta^*} + \frac{1}{\beta} = 1$. Define

$$I(\alpha) = \begin{cases} \sqrt{\min\{m, n\}}, & \alpha = \infty, \\ 1, & \text{otherwise,} \end{cases} \quad J(\beta) = \begin{cases} \sqrt{p}, & \beta = \infty, \\ 1, & \text{otherwise.} \end{cases}$$

Then, it is easy to show that

$$\frac{1}{I(\alpha^*)} \|\sigma(X)\|_{\alpha} \leq \|X\|_F \leq I(\alpha) \|\sigma(X)\|_{\alpha}, \quad \frac{1}{J(\beta^*)} \|x\|_{\beta} \leq \|x\|_2 \leq J(\beta) \|x\|_{\beta}. \quad (2.4)$$

THEOREM 2.1. *Let $\mathcal{X} = \{X^{(k)} : k \in \mathbb{Z}_+\}$ denote the sequence of iterates generated by the First-Order Augmented Lagrangian Algorithm (FALC) displayed in Figure 2.1 for a fixed sequence of parameters $\{\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)}\}_{k \in \mathbb{Z}}$ such that*

- (i) *penalty multipliers, $\lambda^{(k)} \searrow 0$, and*
 - (ii) *approximate optimality parameters, $\epsilon^{(k)} \searrow 0$ such that $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$ for all $k \geq 1$ for some $B > 0$.*
 - (iii) *subgradient tolerance parameters, $\tau^{(k)} \searrow 0$ and $\xi^{(k)} \searrow 0$ such that $\frac{\tau^{(k)}}{\lambda^{(k)}} \rightarrow 0$ and $\frac{\xi^{(k)}}{\lambda^{(k)}} \rightarrow 0$ as $k \rightarrow \infty$.*
- Then $\mathcal{X} = \{X^{(k)} : k \in \mathbb{Z}_+\}$ is a bounded sequence and any limit point \bar{X} of this sequence $\{X^{(k)}\}_{k \in \mathbb{Z}_+}$ is an optimal solution of the composite norm minimization problem (1.1).*

Remark 2.1. *The notation $\gamma^{(k)} \searrow \eta$ (resp. $\gamma^{(k)} \nearrow \eta$) denotes that the sequence $\{\gamma^{(k)}\}_{k \in \mathbb{Z}_+}$ is monotonically decreasing (resp. increasing).*

Proof. Due to Lemma 2.3, we are guaranteed that each inner loop terminates after a finite number of steps. Hence, $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k \in \mathbb{Z}_+}$ sequence exists.

First, we show that for all $k > 1$,

$$\|\theta_2^{(k)}\|_2 \leq \max \left\{ \sigma_{\max}(M) \sqrt{\frac{2\epsilon^{(k-1)}}{(\lambda^{(k-1)})^2}}, \frac{\tau^{(k-1)}}{\lambda^{(k-1)}} \right\} + J(\beta^*) \mu_2, \quad (2.5a)$$

$$\|\mathcal{A}^*(\theta_1^{(k)})\|_F \leq \|\mathcal{C}^*(\theta_2^{(k)})\|_F + \max \left\{ \sigma_{\max}(M) \sqrt{\frac{2\epsilon^{(k-1)}}{(\lambda^{(k-1)})^2}}, \frac{\tau^{(k-1)}}{\lambda^{(k-1)}} \right\} + I(\alpha^*) \mu_1. \quad (2.5b)$$

Consider the following two cases.

- (a) the k -th inner loop terminates with the iterate $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfying (2.3)(a). Then Corollary A.2 guarantees that

$$\|\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)} \theta_2^{(k)}\|_2 \leq \sqrt{2\epsilon^{(k)}} \sigma_{\max}(M) + J(\beta^*) \lambda^{(k)} \mu_2, \quad (2.6a)$$

$$\begin{aligned} \|\mathcal{A}^*(\mathcal{A}(X^{(k)}) + y^{(k)} - b - \lambda^{(k)} \theta_1^{(k)}) + \mathcal{C}^*(\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)} \theta_2^{(k)})\|_F \\ \leq \sqrt{2\epsilon^{(k)}} \sigma_{\max}(M) + I(\alpha^*) \lambda^{(k)} \mu_1. \end{aligned} \quad (2.6b)$$

- (b) the k -th inner loop terminates with an iterate $(X^{(k)}, s^{(k)}, y^{(k)})$ that satisfies (2.3)(b). Hence, there exists $Q^{(k)} \in \partial \|\sigma(\cdot)\|_{\alpha} |_{X^{(k)}}$ and $q^{(k)} \in \partial \|\cdot\|_{\beta} |_{s^{(k)}}$ such that

$$\sqrt{\|\lambda^{(k)} \mu_1 Q^{(k)} + \nabla_X f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_F^2 + \|\lambda^{(k)} \mu_2 q^{(k)} + \nabla_s f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2^2} \leq \tau^{(k)}.$$

Since $\|q^{(k)}\|_{\beta^*} \leq 1$ and $\|\sigma(Q^{(k)})\|_{\alpha^*} \leq 1$, from the definition of $I(\cdot)$ and $J(\cdot)$ in (2.4), it follows that $\|\sigma(Q^{(k)})\|_F \leq I(\alpha^*)$ and $\|q^{(k)}\|_2 \leq J(\beta^*)$. Then we have

$$\|\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)}\|_2 \leq \tau^{(k)} + J(\beta^*)\lambda^{(k)}\mu_2, \quad (2.7a)$$

$$\begin{aligned} \|\mathcal{A}^*(\mathcal{A}(X^{(k)}) + y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)}) + \mathcal{C}^*(\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)})\|_F \\ \leq \tau^{(k)} + I(\alpha^*)\lambda^{(k)}\mu_1. \end{aligned} \quad (2.7b)$$

Thus, from (2.6) and (2.7), it follows that for all $k \geq 1$

$$\|\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)}\|_2 \leq \max\left\{\sqrt{2\epsilon^{(k)}} \sigma_{\max}(M), \tau^{(k)}\right\} + J(\beta^*)\lambda^{(k)}\mu_2, \quad (2.8a)$$

$$\begin{aligned} \|\mathcal{A}^*(\mathcal{A}(X^{(k)}) + y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)})\|_F \leq \max\left\{\sqrt{2\epsilon^{(k)}} \sigma_{\max}(M), \tau^{(k)}\right\} \\ + \|\mathcal{C}^*(\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)})\|_F + I(\alpha^*)\lambda^{(k)}\mu_1. \end{aligned} \quad (2.8b)$$

Since $\theta_1^{(k+1)} = \theta_1^{(k)} - \frac{\mathcal{A}(X^{(k)}) + y^{(k)} - b}{\lambda^{(k)}}$ and $\theta_2^{(k+1)} = \theta_2^{(k)} - \frac{\mathcal{C}(X^{(k)}) + s^{(k)} - d}{\lambda^{(k)}}$, (2.5) follows from (2.6) and (2.7). Thus, $\{(\theta_1^{(k)}, \theta_2^{(k)})\}_{k \in \mathbb{Z}}$ satisfies (2.5). Because A has full row-rank, $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B_1$ and $\frac{\tau^{(k)}}{\lambda^{(k)}} \rightarrow 0$, (2.5) implies that there exist $B_{\theta_1} > 0$ and $B_{\theta_2} > 0$ such that for all $k \geq 1$

$$\|\theta_1^{(k)}\|_2 \leq B_{\theta_1}, \quad \|\theta_2^{(k)}\|_2 \leq B_{\theta_2}. \quad (2.9)$$

From (2.9), it follows that for $i = 1, 2$,

$$\lim_{k \rightarrow \infty} \lambda^{(k)}\theta_i^{(k)} = 0, \quad (2.10)$$

and

$$\lim_{k \rightarrow \infty} \lambda^{(k)}\|\theta_i^{(k)}\|_2^2 = 0. \quad (2.11)$$

Also, $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$ for all $k \geq 1$ implies that

$$\lim_{k \rightarrow \infty} \frac{\epsilon^{(k)}}{\lambda^{(k)}} = 0. \quad (2.12)$$

We next show that $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k \in \mathbb{Z}_+}$ is a bounded sequence. Consider the following two possibilities.

(a) $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfies (2.3)(a). Recall that $X^{(0)} = \operatorname{argmin}\{\|X\|_F : \mathcal{A}(X) = b\}$. Define $s^{(0)} = d - \mathcal{C}(X^{(0)})$ and $y^{(0)} = b - \mathcal{A}(X^{(0)}) = 0$. Then

$$P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \leq P^{(k)}(X^{(0)}, s^{(0)}, y^{(0)}) = \lambda^{(k)}\eta + \frac{1}{2} \left(\|\lambda^{(k)}\theta_1^{(k)}\|_2^2 + \|\lambda^{(k)}\theta_2^{(k)}\|_2^2 \right),$$

where $\eta := \mu_1\|\sigma(X^{(0)})\|_\alpha + \mu_2\|\mathcal{C}(X^{(0)}) - d\|_\beta$. Hence,

$$\mu_1\|\sigma(X^{(k)})\|_\alpha \leq \frac{P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})}{\lambda^{(k)}} \leq \frac{P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) + \epsilon^{(k)}}{\lambda^{(k)}} \leq \eta^{(k)} + \frac{\epsilon^{(k)}}{\lambda^{(k)}}, \quad (2.13)$$

where $\eta^{(k)} = \eta + \frac{\lambda^{(k)}}{2}\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2$.

(b) $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfies (2.3)(b). Then trivially, $\mu_1\|\sigma(X^{(k)})\|_\alpha \leq \eta^{(k)}$. Hence, from (2.13), we can conclude that for all $k \geq 1$, $\mu_1\|\sigma(X^{(k)})\|_\alpha \leq \eta^{(k)} + \frac{\epsilon^{(k)}}{\lambda^{(k)}}$. Hence, $\mu_1\|\sigma(X^{(k)})\|_\alpha \leq \eta + \lambda^{(k)} \left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2} + B \right)$ for all $k \geq 1$.

Therefore, we can conclude that there exists a subsequence $\mathcal{K} \subset \mathbb{Z}_+$ such that $\lim_{k \in \mathcal{K}} X^{(k)} = \bar{X}$ exists. Furthermore, (2.6a) and (2.7a) guarantee that $\lim_{k \in \mathcal{K}} s^{(k)} = \bar{s}$ exists and similarly (2.6b) and (2.7b) guarantee that $\lim_{k \in \mathcal{K}} y^{(k)} = \bar{y}$ exists. In the rest of the proof, we will show that $\bar{X} \in \operatorname{argmin}\{\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta : \|\mathcal{A}(X) - b\|_\gamma \leq \rho\}$. We consider the following two cases.

- (a) There exists a further subsequence $\mathcal{K}_1 \subset \mathcal{K}$ such that for all $k \in \mathcal{K}_1$, $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfies (2.3)(a), i.e. the sequence $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k \in \mathbb{Z}_+}$ computed in Step 1 of FALC satisfies

$$0 \leq P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) - P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \leq \epsilon^{(k)} \quad \forall k \geq 1. \quad (2.14)$$

Let X_* denote any optimal solution of the composite norm minimization problem and let $s_* = d - \mathcal{C}(X_*)$ and $y_* = b - \mathcal{A}(X_*)$. Since $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) = \operatorname{argmin}\{P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$, for $k \geq 1$, it follows that $P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \leq P^{(k)}(X_*, s_*, y_*)$. Thus, (2.14) implies that $P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) \leq P^{(k)}(X_*, s_*, y_*) + \epsilon^{(k)}$. Hence, for all $k \geq 1$,

$$\begin{aligned} \mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta &\leq \frac{P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})}{\lambda^{(k)}} \\ &\leq \frac{P^{(k)}(X_*, s_*, y_*) + \epsilon^{(k)}}{\lambda^{(k)}} \\ &= \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2 \right) + \frac{\epsilon^{(k)}}{\lambda^{(k)}}. \end{aligned} \quad (2.15)$$

Taking the limit of both sides of (2.15) along the subsequence \mathcal{K}_1 , and using the fact that $\bar{s} = d - \mathcal{C}(\bar{X})$, we get

$$\begin{aligned} \mu_1 \|\sigma(\bar{X})\|_\alpha + \mu_2 \|\mathcal{C}(\bar{X}) - d\|_\beta &= \lim_{k \in \mathcal{K}_1} \mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta \\ &\leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \lim_{k \in \mathcal{K}_1} \left\{ \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2 \right) + \frac{\epsilon^{(k)}}{\lambda^{(k)}} \right\} \\ &= \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta, \end{aligned} \quad (2.16)$$

where (2.16) follows from the fact that $\{\theta_i^{(k)}\}$ is uniformly bounded for $i = 1, 2$, $\lambda^{(k)} \rightarrow 0$, and $\epsilon^{(k)}/\lambda^{(k)} \rightarrow 0$. Taking the limit of both sides of (2.8b) of along \mathcal{K}_1 and using (2.10), we get

$$\|\mathcal{A}^*(\mathcal{A}(\bar{X}) + \bar{y} - b)\|_F \leq 0, \quad (2.17)$$

and since A has full row rank, it follows that $\mathcal{A}(\bar{X}) + \bar{y} = b$. Since $\|y^{(k)}\|_\gamma \leq \rho$ for all $k \geq 1$, we can conclude that \bar{X} is feasible, i.e. $\|\mathcal{A}(\bar{X}) - b\|_\gamma \leq \rho$. Thus, from (2.16) and the fact that $X_* \in \operatorname{argmin}\{\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta : \|\mathcal{A}(X) - b\|_\gamma \leq \rho\}$, it follows that \bar{X} is an optimal solution for the composite norm minimization problem (1.1).

- (b) There exists $K \in \mathcal{K}$ such that, for all $k \in \mathcal{K}_2 = \mathcal{K} \cap \{k \geq K\}$, the inner iterations for the k -th subproblem terminates with an iterate $(X^{(k)}, s^{(k)}, y^{(k)})$ that satisfies (2.3)(b).

For all $k \in \mathcal{K}_2$, there exist $Q^{(k)} \in \partial \|\sigma(\cdot)\|_\alpha|_{X^{(k)}}$ and $q^{(k)} \in \partial \|\cdot\|_\beta|_{s^{(k)}}$ such that (2.3)(b) holds. Hence, we have

$$\|\lambda^{(k)} \mu_2 q^{(k)} + \nabla_s f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2 \leq \tau^{(k)}, \quad (2.18a)$$

$$\|\lambda^{(k)} \mu_1 Q^{(k)} + \nabla_X f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_F \leq \tau^{(k)}, \quad (2.18b)$$

$$\rho \|\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_{\gamma^*} + \nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T y^{(k)} \leq \xi^{(k)}. \quad (2.18c)$$

Since $\|\theta_i^{(k)}\|_2 \leq B_{\theta_i}$, for all $k \geq 1$, and $\lim_{k \rightarrow \infty} \lambda^{(k)} \theta_i^{(k)} = 0$ for $i \in \{1, 2\}$. Taking the limit of both sides of (2.7a) for $k \in \mathcal{K}_2$, we have $\|\mathcal{C}(\bar{X}) + \bar{s} - d\|_2 \leq 0$, i.e. $\bar{s} = d - \mathcal{C}(\bar{X})$. Moreover, taking the limit of both sides of (2.7b) for $k \in \mathcal{K}$ and using the fact that $\bar{s} = d - \mathcal{C}(\bar{X})$, we have $\|\mathcal{A}^*(\mathcal{A}(\bar{X}) + \bar{y} - b)\|_2 \leq 0$. Since A as full row rank, it follows that

$$\mathcal{A}(\bar{X}) + \bar{y} = b, \quad \|\bar{y}\|_\gamma \leq \rho. \quad (2.19)$$

For all $k \in \mathcal{K}_2$, $Q^{(k)} \in \partial\|\sigma(\cdot)\|_\alpha|_{X^{(k)}}$ and $q^{(k)} \in \partial\|\cdot\|_\beta|_{s^{(k)}}$, therefore, $\|\sigma(Q^{(k)})\|_{\alpha^*} \leq 1$ and $\|q^{(k)}\|_{\beta^*} \leq 1$. Hence, there exists a subsequence $\mathcal{K}_3 \subset \mathcal{K}_2$ such that $\lim_{k \in \mathcal{K}_3} (Q^{(k)}, q^{(k)}) = (\bar{Q}, \bar{q})$ exists. One can easily show that $\bar{Q} \in \partial\|\sigma(\cdot)\|_\alpha|_{\bar{X}}$ and $\bar{q} \in \partial\|\cdot\|_\beta|_{\bar{s}}$. Dividing both sides of (2.18a) by $\lambda^{(k)}$, we get

$$\|\mu_2 Q^{(k)} - \theta_2^{(k+1)}\|_2 \leq \frac{\tau^{(k)}}{\lambda^{(k)}}, \quad (2.20)$$

for all $k \in \mathcal{K}_2 \supset \mathcal{K}_3$. Since $\lim_{k \in \mathcal{K}_3} q^{(k)} = \bar{q}$ and $\lim_{k \in \mathbb{Z}_+} \frac{\tau^{(k)}}{\lambda^{(k)}} = 0$, it follows that $\lim_{k \in \mathcal{K}_3} \theta_2^{(k+1)} = \bar{\theta}_2$ exists and taking the limit of both sides of (2.20), we have

$$\mu_2 \bar{q} = \bar{\theta}_2.$$

Dividing both sides of (2.18b) by $\lambda^{(k)}$, we get

$$\|\mu_1 Q^{(k)} - \mathcal{A}^*(\theta_1^{(k+1)}) - \mathcal{C}^*(\theta_2^{(k+1)})\|_F \leq \frac{\tau^{(k)}}{\lambda^{(k)}}, \quad (2.21)$$

for all $k \in \mathcal{K}_2 \supset \mathcal{K}_3$. Since $\lim_{k \in \mathcal{K}_3} Q^{(k)} = \bar{Q}$, $\lim_{k \in \mathbb{Z}_+} \frac{\tau^{(k)}}{\lambda^{(k)}} = 0$ and A has full row rank, it follows that $\lim_{k \in \mathcal{K}_3} \theta_1^{(k+1)} = \bar{\theta}_1$ exists and taking the limit of both sides of (2.21), we have $\mu_1 \bar{Q} - \mu_2 \mathcal{C}^*(\bar{q}) = \mathcal{A}^*(\bar{\theta}_1)$. Note that $\bar{q} \in \partial\|\cdot\|_\beta|_{\bar{s}}$ and $\bar{s} = d - \mathcal{C}(\bar{X})$. Hence, $-\mathcal{C}^*(\bar{q}) \in \partial\|d - \mathcal{C}(\cdot)\|_\beta|_{\bar{X}}$ and we have

$$\mathcal{A}^*(\bar{\theta}_1) = G_* \quad G_* \in \partial\mu_1\|\sigma(\cdot)\|_\alpha + \mu_2\|d - \mathcal{C}(\cdot)\|_\beta|_{\bar{X}}, \quad (2.22)$$

where $G_* := \mu_1 \bar{Q} - \mu_2 \mathcal{C}^*(\bar{q})$. Dividing both sides of (2.18c) by $\lambda^{(k)}$, we get

$$\rho\|\theta_1^{(k+1)}\|_{\gamma^*} - \left(\theta_1^{(k+1)}\right)^T y^{(k)} \leq \frac{\tau^{(k)}}{\lambda^{(k)}}, \quad (2.23)$$

for all $k \in \mathcal{K}_2 \supset \mathcal{K}_3$. Since $\lim_{k \in \mathcal{K}_3} \theta_1^{(k+1)} = \bar{\theta}_1$, taking the limit of both sides of (2.23) and multiplying by -1 , we have

$$0 \leq -\rho\|\bar{\theta}_1\|_{\gamma^*} + (\bar{\theta}_1)^T \bar{y} = \min_{y: \|y\|_\gamma \leq \rho} -(\bar{\theta}_1)^T (y - \bar{y}).$$

Thus,

$$-(\bar{\theta}_1)^T (y - \bar{y}) \geq 0 \quad \forall y: \|y\|_\gamma \leq \rho. \quad (2.24)$$

Consequently, (2.22) and (2.24) together imply that (\bar{X}, \bar{y}) satisfies the first order optimality conditions of the relaxed problem (2.25).

$$\min_{X \in \mathbb{R}^{m \times n}, y \in \mathbb{R}^q} \left\{ \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta - (\bar{\theta}_1)^T (\mathcal{A}(X) + y - b) : \|y\|_\gamma \leq \rho \right\}. \quad (2.25)$$

Since (2.25) is convex, it follows that (\bar{X}, \bar{y}) is an optimal solution to the relaxed problem (2.25). Moreover, from (2.19), (\bar{X}, \bar{y}) is feasible to the composite norm minimization problem, i.e. $\min\{\mu_1 \|X\|_\alpha + \mu_2 \|d - \mathcal{C}(X)\|_\beta : \mathcal{A}(X) + y = b, \|y\|_\gamma \leq \rho\}$. Therefore, $\bar{X} \in \operatorname{argmin}_{X \in \mathbb{R}^{m \times n}} \{\mu_1 \|X\|_\alpha + \mu_2 \|d - \mathcal{C}(X)\|_\beta : \|\mathcal{A}(X) - b\|_\gamma \leq \rho\}$.

□

For compressed sensing and matrix completion problems exact recovery occurs only when $\min_{x \in \mathbb{R}^n} \{\|x\|_1 : Ax = b\}$ and $\min_{X \in \mathbb{R}^{m \times n}} \{\|X\|_* : X_{ij} = M_{ij} \ (i, j) \in \Omega\}$ have both *unique* solutions, respectively. The following Corollary establishes that FALC converges to this solution.

COROLLARY 2.2. *Suppose the composite norm minimization problem (1.1) has a unique optimal solution X_* . Let $\{X^{(k)} : k \in \mathbb{Z}_+\}$ denote the sequence of iterates generated by the First-Order Augmented Lagrangian Algorithm (FALC) displayed in Figure 2.1 when the sequence of $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k \in \mathbb{Z}}$ satisfies all the conditions in (2.1). Then $\lim_{k \rightarrow \infty} X^{(k)} = X_*$ where $X_* = \operatorname{argmin}_{X \in \mathbb{R}^{m \times n}} \{\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta : \|\mathcal{A}(X) - b\|_\gamma \leq \rho\}$.*

We next establish a bound on the complexity of computing the iterate $(X^{(k)}, s^{(k)}, y^{(k)})$ using Algorithm 3 in [33].

LEMMA 2.3. *For all $k \geq 1$, the worst-case complexity of computing $(X^{(k)}, s^{(k)}, y^{(k)})$ is*

$$\mathcal{O}\left(\frac{\min\{nm^2, n^2m\}}{\sqrt{\epsilon^{(k)}}}\right), \quad (2.26)$$

when $\lambda^{(k)} \rightarrow 0$, $\epsilon^{(k)} \rightarrow 0$ such that $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$ for all $k \geq 1$.

Proof. Let $\{(X^{(k,\ell)}, s^{(k,\ell)}, y^{(k,\ell)})\}_{\ell \in \mathbb{Z}_+}$ denote the iterates computed by Algorithm 3 in [33] when applied to the k -th sub-problem

$$\min \left\{ P^{(k)}(X, s, y) = \lambda^{(k)}(\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho \right\},$$

with the prox function $h^{(k)}(X, s, y) = \frac{1}{2}\|X - X^{(k-1)}\|_F^2 + \frac{1}{2}\|s - s^{(k-1)}\|_2^2 + \frac{1}{2}\|y - y^{(k-1)}\|_2^2$ and initial iterate $(X^{(k-1)}, s^{(k-1)}, y^{(k-1)}) \in \mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q$. Then Corollary 3 in [33] establishes that for all iterates $\ell \geq \sqrt{\frac{4Lh^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})}{\epsilon^{(k)}}} - 1$,

$$P^{(k)}(X^{(k,\ell)}, s^{(k,\ell)}, y^{(k,\ell)}) \leq \inf \{ P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho \} + \epsilon^{(k)},$$

where $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) = \operatorname{argmin}\{P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$, $L = \sigma_{max}^2 \begin{pmatrix} I & 0 & C \\ 0 & I & A \end{pmatrix}$ is the Lipschitz constant of $\nabla f^{(k)}$ for all $k \geq 1$.

Let X_* denote the optimal solution of (1.1), $s_* = d - \mathcal{C}(X_*)$ and $y_* = b - \mathcal{A}(X_*)$. Then we have $\|y_*\|_\gamma \leq \rho$. Since (X_*, s_*, y_*) is feasible to the k -th subproblem, we have $P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \leq P^{(k)}(X_*, s_*, y_*)$, which implies

$$\mu_1 \|\sigma(X_*^{(k)})\|_\alpha + \mu_2 \|s_*^{(k)}\|_\beta \leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2 \right).$$

From (2.15), it follows that the inexact minimizer of $(k-1)$ -th subproblem $(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})$ satisfies

$$\mu_1 \|\sigma(X^{(k-1)})\|_\alpha + \mu_2 \|s^{(k-1)}\|_\beta \leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k-1)}}{2} \left(\|\theta_1^{(k-1)}\|_2^2 + \|\theta_2^{(k-1)}\|_2^2 \right) + \frac{\epsilon^{(k-1)}}{\lambda^{(k-1)}}.$$

Since $\theta_1^{(k)} \leq B_{\theta_1}$ and $\theta_2^{(k)} \leq B_{\theta_2}$ for all $k \geq 1$ —see (2.9), and $\{\lambda^{(k)}\}_{k \in \mathbb{Z}_+}$ is a decreasing sequence, it follows that

$$\begin{aligned} & \mu_1 \left(\|\sigma(X^{(k-1)})\|_\alpha + \|\sigma(X_*^{(k)})\|_\alpha \right) + \mu_2 \left(\|s^{(k-1)}\|_\beta + \|s_*^{(k)}\|_\beta \right) \\ & \leq 2 \left(\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta \right) + \lambda^{(k-1)} (B_{\theta_1}^2 + B_{\theta_2}^2) + \frac{\epsilon^{(k-1)}}{\lambda^{(k-1)}}. \end{aligned} \quad (2.27)$$

From the definition of $h^{(k)}(\cdot)$ it follows that

$$\begin{aligned} & h^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \\ & = \frac{1}{2}\|X_*^{(k)} - X^{(k-1)}\|_F^2 + \frac{1}{2}\|s_*^{(k)} - s^{(k-1)}\|_2^2 + \frac{1}{2}\|y_*^{(k)} - y^{(k-1)}\|_2^2, \\ & \leq \frac{1}{2}I^2(\alpha) \left(\|\sigma(X_*^{(k)})\|_\alpha + \|\sigma(X^{(k-1)})\|_\alpha \right)^2 + \frac{1}{2}J^2(\beta) \left(\|s_*^{(k)}\|_\beta + \|s^{(k-1)}\|_\beta \right)^2 + \frac{1}{2}J^2(\gamma) \left(\|y_*^{(k)}\|_\gamma + \|y^{(k-1)}\|_\gamma \right)^2, \\ & \leq \frac{1}{2} \left(\frac{I^2(\alpha)}{\mu_1^2} \frac{J^2(\beta)}{\mu_2^2} \right) \left(2(\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta) + \lambda^{(k-1)} (B_{\theta_1}^2 + B_{\theta_2}^2) + \frac{\epsilon^{(k-1)}}{\lambda^{(k-1)}} \right)^2 \\ & \quad + 2\rho^2 J^2(\gamma), \end{aligned} \quad (2.28)$$

where (2.28) follows from the fact that $\|y_*^{(k)}\| \leq \rho$, $\|y^{(k-1)}\| \leq \rho$ and (2.27) together with the definition of $I(\cdot)$ and $J(\cdot)$ in (2.4). Thus,

$$\sqrt{\frac{4Lh^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})}{\epsilon^{(k)}}} \leq \frac{C\sigma_{\max}(M)}{\epsilon^{(k)}},$$

where

$$C = \sqrt{8 \left(\frac{I^2(\alpha)}{\mu_1^2} + \frac{J^2(\beta)}{\mu_2^2} \right) \left(\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(1)}}{2} (B_{\theta_1}^2 + B_{\theta_1}^2 + B) \right) + \rho^2 J^2(\gamma)}.$$

Each step of Algorithm 3 in [33] involves computing the solution of the following optimization problems.

(a) one *matrix shrinkage problem* of the form

$$\min_{X \in \mathbb{R}^{m \times n}} \left\{ \delta_1 \|\sigma(X)\|_\alpha + \frac{1}{2} \|X - Y\|_F^2 : \|\sigma(X)\|_\alpha \leq \rho_1 \right\} \quad (2.29)$$

Lemma A.4 establishes when $\alpha \in \{1, \infty\}$ the worst-case complexity of computing a solution to (2.29) is the same that of computing a full SVD, i.e. $\mathcal{O}(\min\{nm^2, n^2m\})$, and when $\alpha = 2$, the worst-case complexity is $\mathcal{O}(mn)$.

(b) one *vector shrinkage problem* of the form

$$\min_{s \in \mathbb{R}^p} \left\{ \delta_2 \|s\|_\beta + \frac{1}{2} \|s - q\|_2^2 : \|s\|_\beta \leq \rho_2 \right\} \quad (2.30)$$

for a given $q \in \mathbb{R}^p$. Lemma A.4 establishes that the complexity of solving the vector shrinkage problem is $\mathcal{O}(p \log(p))$ when $\beta \in \{1, \infty\}$ and $\mathcal{O}(p)$ when $\beta = 2$.

(c) one *vector shrinkage problem* of the form

$$\min_{y \in \mathbb{R}^q} \left\{ \frac{1}{2} \|y - z\|_2^2 : \|y\|_\gamma \leq \rho_3 \right\} \quad (2.31)$$

for a given $z \in \mathbb{R}^q$. Lemma A.4 establishes that the complexity of solving the vector shrinkage problem is $\mathcal{O}(q \log(q))$ when $\gamma \in \{1, \infty\}$ and $\mathcal{O}(q)$ when $\gamma = 2$.

□

Next, we characterize the finite iteration performance of FALC. This analysis will lead to a convergence rate result in Theorem 2.5.

THEOREM 2.4. *Let $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k \in \mathbb{Z}_+}$ denote the sequence of iterates generated by the FALC displayed in Figure 2.1. Suppose there exists $B > 0$ such that $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$, $\tau^{(k)} = \kappa_1 \epsilon^{(k)}$ and $\xi^{(k)} = \kappa_2 \epsilon^{(k)}$ for all $k \geq 1$ so that $\lim_{k \rightarrow \infty} \frac{\tau^{(k)}}{\lambda^{(k)}} \rightarrow 0$ and $\lim_{k \rightarrow \infty} \frac{\xi^{(k)}}{\lambda^{(k)}} \rightarrow 0$ as $k \rightarrow \infty$. Then there exists $c_1 > 0$, $c_2 > 0$ and $c_3 > 0$ such that for all $k \geq 1$,*

$$(i) \quad \|y^{(k)}\|_\gamma \leq \rho \text{ such that } \|\mathcal{A}(X^{(k)}) + y^{(k)} - b\|_2 \leq c_1 \lambda^{(k)},$$

$$(ii) \quad |(\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta) - (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta)| \leq c_2 \lambda^{(k)} + c_3 \sqrt{\epsilon^{(k)}},$$

where $X_* = \operatorname{argmin}_{X \in \mathbb{R}^{m \times n}} \{\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta : \mathcal{A}(X) = b\}$.

Proof. Given $k \geq 1$, consider the following two cases:

(a) $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfies (2.3)(a). Then, from (2.15) it follows that

$$\begin{aligned} \mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta &\leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta \\ &\quad + \frac{\lambda^{(k)}}{2} (\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2) + \frac{\epsilon^{(k)}}{\lambda^{(k)}}. \end{aligned} \quad (2.32)$$

(b) $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfies (2.3)(b). Then from the convexity of $P^{(k)}$, it follows that

$$\begin{aligned}
& P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) \\
& \leq P^{(k)}(X_*, s_*) - \left\langle G, X_* - X^{(k)} \right\rangle - g^T(s_* - s^{(k)}) - \nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T (y_* - y^{(k)}), \\
& \leq P^{(k)}(X_*, s_*) + \|G\|_F \|X_* - X^{(k)}\|_F + \|g\|_2 \|s_* - s^{(k)}\|_2 - \min_{y: \|y\|_\gamma \leq \rho} \nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T (y - y^{(k)}), \\
& \leq P^{(k)}(X_*, s_*) + \tau^{(k)}(\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2) + \nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T y^{(k)} \\
& \quad + \rho \|\nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_{\gamma^*}, \\
& \leq P^{(k)}(X_*, s_*) + \tau^{(k)}(\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2) + \xi^{(k)}
\end{aligned} \tag{2.33}$$

where $(G, g) \in \partial P^{(k)}(.,.)|_{(X^{(k)}, s^{(k)})}$ and $\partial P^{(k)}(.,.)|_{(X^{(k)}, s^{(k)})}$ denotes the set of subgradients of the function $P^{(k)}$ at $(X^{(k)}, s^{(k)})$. Hence, it follows that

$$\begin{aligned}
\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta & \leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2 \right) \\
& \quad + \frac{\tau^{(k)}}{\lambda^{(k)}} (\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2) + \frac{\xi^{(k)}}{\lambda^{(k)}},
\end{aligned} \tag{2.34}$$

Thus, from (2.32) and (2.34), it follows that for all $k \geq 1$

$$\begin{aligned}
\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta & \leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta + \left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2} \right) \lambda^{(k)} \\
& \quad + \max \left\{ \frac{\epsilon^{(k)}}{\lambda^{(k)}}, \frac{\tau^{(k)}}{\lambda^{(k)}} (\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2) + \frac{\xi^{(k)}}{\lambda^{(k)}} \right\},
\end{aligned} \tag{2.35}$$

where (2.35) follows from the bound on $\|\theta_i^{(k)}\|_2$ for $i \in \{1, 2\}$ established in (2.9). From (2.4) and Corollary A.2 in Appendix A, it follows that for all $k \geq 1$,

$$\begin{aligned}
\|\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)} \theta_2^{(k)}\|_\beta & \leq J(\beta^*) \|\mathcal{C}(X^{(k)}) + s^{(k)} - d - \lambda^{(k)} \theta_2^{(k)}\|_2 \\
& \leq J(\beta^*) \left(\sqrt{2\epsilon^{(k)}} \sigma_{max}(M) + J(\beta^*) \mu_2 \lambda^{(k)} \right).
\end{aligned} \tag{2.36}$$

Triangular inequality and the uniform bound $\|\theta_2^{(k)}\|_2 \leq B_{\theta_2}$, for all $k \geq 1$, established in (2.9), together imply that

$$\begin{aligned}
\|\mathcal{C}(X^{(k)}) - d\|_\beta & \leq \|s^{(k)}\|_\beta + \|\lambda^{(k)} \theta_2^{(k)}\|_\beta + J(\beta^*) \left(\sqrt{2\epsilon^{(k)}} \sigma_{max}(M) + J(\beta^*) \mu_2 \lambda^{(k)} \right), \\
& \leq \|s^{(k)}\|_\beta + J(\beta^*) (B_{\theta_2} + \mu_2 J(\beta^*)) \lambda^{(k)} + \left(\sqrt{2} \sigma_{max}(M) J(\beta^*) \right) \sqrt{\epsilon^{(k)}}.
\end{aligned} \tag{2.37}$$

Thus, (2.35) and (2.37) together imply that

$$\begin{aligned}
\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta & \leq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta \\
& \quad + \left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2} + \mu_2 J(\beta^*) (B_{\theta_2} + \mu_2 J(\beta^*)) \right) \lambda^{(k)} \\
& \quad + \max \left\{ \frac{\epsilon^{(k)}}{\lambda^{(k)}}, \frac{\tau^{(k)}}{\lambda^{(k)}} (\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2) + \frac{\xi^{(k)}}{\lambda^{(k)}} \right\} \\
& \quad + \mu_2 \left(\sqrt{2} \sigma_{max}(M) J(\beta^*) \right) \sqrt{\epsilon^{(k)}}.
\end{aligned} \tag{2.38}$$

Since $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k \in \mathbb{Z}_+}$ is a bounded sequence, $\tau^{(k)} = \kappa_1 \epsilon^{(k)}$ and $\xi^{(k)} = \kappa_2 \epsilon^{(k)}$ for all $k \geq 1$, (2.38) implies one side of the bound in (ii).

For all $k \geq 1$,

$$\begin{aligned}\|\mathcal{A}(X^{(k)}) + y^{(k)} - b\|_2 &\leq \|\mathcal{A}(X^{(k)}) + y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)}\|_2 + \lambda^{(k)}\|\theta_1^{(k)}\|_2, \\ &= \lambda^{(k)}\|\theta_1^{(k+1)}\|_2 + \lambda^{(k)}\|\theta_1^{(k)}\|_2, \\ &\leq 2B_{\theta_1}\lambda^{(k)},\end{aligned}$$

where the last inequality follows the fact that $\|\theta_1^{(k)}\|_2 \leq B_{\theta_1}$ for all $k \geq 1$; see (2.9) for details. This establishes (i).

Next, we establish a lower bound for $P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})$ using the following pair of Lagrangian duals

$$\begin{aligned}\min_{X \in \mathbb{R}^{m \times n}} \quad & \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta, \\ \text{s.t.} \quad & \|\mathcal{A}(X) - b\|_\gamma \leq \rho.\end{aligned}\tag{2.39a}$$

$$\begin{aligned}\max_{w \in \mathbb{R}^q, v \in \mathbb{R}^p} \quad & b^T w + d^T v - \rho \|w\|_{\gamma^*}, \\ \text{s.t.} \quad & \|\sigma(\mathcal{A}^*(w) + \mathcal{C}^*(v))\|_{\alpha^*} \leq \mu_1, \\ & \|v\|_{\beta^*} \leq \mu_2.\end{aligned}\tag{2.39b}$$

Let (w_*, v_*) denote the optimal solution of the dual (2.39b). Let $f(X, s, y) = \frac{1}{2}\|\mathcal{A}(X) + y - b - \lambda\theta_1\|_2^2 + \frac{1}{2}\|\mathcal{C}(X) + s - d - \lambda\theta_2\|_2^2$. Moreover, (2.40a) and (2.40b) below are also a Lagrange primal-dual pair of problems.

$$\begin{aligned}\min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y \in \mathbb{R}^q} \quad & \lambda(\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f(X, s, y), \\ \text{s.t.} \quad & \|y\|_\gamma \leq \rho.\end{aligned}\tag{2.40a}$$

$$\begin{aligned}\max_{w \in \mathbb{R}^q, v \in \mathbb{R}^p} \quad & \lambda(b + \lambda\theta_1)^T w + \lambda(d + \lambda\theta_2)^T v - \frac{\lambda^2}{2}(\|w\|_2^2 + \|v\|_2^2) - \lambda\rho\|w\|_{\gamma^*}, \\ \text{s.t.} \quad & \|\sigma(\mathcal{A}^*(w) + \mathcal{C}^*(v))\|_{\alpha^*} \leq \mu_1, \\ & \|v\|_{\beta^*} \leq \mu_2.\end{aligned}\tag{2.40b}$$

Since (w_*, v_*) is feasible for (2.40b), it follows that

$$\begin{aligned}P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) &= \min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y \in \mathbb{R}^q} \left\{ \lambda^{(k)}(\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f^{(k)}(X, s, y) : \|y\|_\gamma \leq \rho \right\} \\ &\geq \lambda^{(k)} \left(b^T w_* + d^T v_* - \rho \|w_*\|_{\gamma^*} - \frac{\lambda^{(k)}}{2} \left(\|w_*\|_2^2 + \|v_*\|_2^2 - 2(\theta_1^{(k)})^T w_* - 2(\theta_2^{(k)})^T v_* \right) \right),\end{aligned}\tag{2.41}$$

$$\begin{aligned}&\geq \lambda^{(k)} (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta) \\ &\quad - \frac{(\lambda^{(k)})^2}{2} \left(\|w_*\|_2^2 + \|v_*\|_2^2 + 2\|\theta_1^{(k)}\|_2 \|w_*\|_2 + 2\|\theta_2^{(k)}\|_2 \|v_*\|_2 \right),\end{aligned}\tag{2.42}$$

where (2.41) follows from weak duality for primal-dual pair in (2.40), and (2.42) follows from strong duality for primal-dual pair in (2.39) and the Cauchy-Schwartz inequality.

Since the FALC iterates $\{X^{(k)}\}_{k \in \mathbb{Z}}$ satisfy

$$\frac{P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})}{\lambda^{(k)}} = (\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta) + \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k+1)}\|_2^2 + \|\theta_2^{(k+1)}\|_2^2 \right),$$

it follows that

$$\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta \geq \frac{P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})}{\lambda^{(k)}} - \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k+1)}\|_2^2 + \|\theta_2^{(k+1)}\|_2^2 \right).\tag{2.43}$$

Thus, the bound on $\|\theta_i^{(k)}\|_2$, $i \in \{1, 2\}$ established in (2.9), and the inequalities (2.42) and (2.43), together imply that

$$\begin{aligned} \mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|s^{(k)}\|_\beta &\geq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta \\ &\quad - \frac{\lambda^{(k)}}{2} ((B_{\theta_1} + \|w_*\|_2)^2 + (B_{\theta_2} + \|v_*\|_2)^2). \end{aligned} \quad (2.44)$$

The bound $\|\mathcal{A}^*(w_*) + \mathcal{C}^*(v_*)\|_F \leq I(\alpha^*) \|\sigma(\mathcal{A}^*(w_*) + \mathcal{C}^*(v_*))\|_{\alpha^*} \leq I(\alpha^*) \mu_1$ implies that

$$\sigma_{\min}(A) \|w_*\|_2 \leq \|\mathcal{A}^*(w_*)\|_F \leq I(\alpha^*) \mu_1 + \|\mathcal{C}^*(v_*)\|_F \leq I(\alpha^*) \mu_1 + \sigma_{\max}(C) \|v_*\|_2,$$

and the bound $\|v_*\|_{\beta^*} \leq \mu_2$ implies that $\|v_*\|_2 \leq J(\beta^*) \mu_2$.

From (2.36), triangular inequality, and the uniform bound $\|\theta_2^{(k)}\|_2 \leq B_{\theta_2}$, for all $k \geq 1$, established in (2.9), it follows that

$$\|s^{(k)}\|_\beta \leq \|\mathcal{C}(X^{(k)}) - d\|_\beta + J(\beta^*) (B_{\theta_2} + \mu_2 J(\beta^*)) \lambda^{(k)} + \left(\sqrt{2} \sigma_{\max}(M) J(\beta^*) \right) \sqrt{\epsilon^{(k)}}. \quad (2.45)$$

From (2.44) and (2.45), it follows that

$$\begin{aligned} \mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta &\geq \mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta \\ &\quad - \left(\frac{(B_{\theta_1} + \|w_*\|_2)^2 + (B_{\theta_2} + \|v_*\|_2)^2}{2} + \mu_2 J(\beta^*) (B_{\theta_2} + \mu_2 J(\beta^*)) \right) \lambda^{(k)} \\ &\quad - \mu_2 \left(\sqrt{2} \sigma_{\max}(M) J(\beta^*) \right) \sqrt{\epsilon^{(k)}}. \end{aligned}$$

This establishes the result. \square

THEOREM 2.5. Fix $0 < \nu < 1$, and strictly positive parameters $(\lambda^{(1)}, \epsilon^{(1)}, \tau^{(1)}, \xi^{(1)})$. Then there exists a sequence of parameters $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k \in \mathbb{Z}_+}$ such that for all $\epsilon > 0$, Algorithm FALC displayed in Figure 2.1, computes an ϵ -feasible and ϵ -optimal solution $\bar{X} \in \mathbb{R}^{m \times n}$ to problem (1.1), i.e. for some $\bar{y} \in \mathbb{R}^q$ such that $\|\bar{y}\|_\gamma \leq \rho$, we have

$$\|\mathcal{A}(\bar{X}) + \bar{y} - b\|_2 \leq \epsilon, \quad |(\mu_1 \|\sigma(\bar{X})\|_\alpha + \mu_2 \|\mathcal{C}(\bar{X}) - d\|_\beta) - (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta)| \leq \epsilon,$$

in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ operations.

Proof. Fix $\lambda^{(1)} > 0$, $\epsilon^{(1)} > 0$ and choose $0 < \nu < 1$ and update the parameters as follows: for all $k \geq 1$,

$$\begin{aligned} \lambda^{(k+1)} &= \nu \lambda^{(k)}, & \xi^{(k)} &= \frac{1}{2} \epsilon^{(k)}, \\ \epsilon^{(k+1)} &= \nu^2 \epsilon^{(k)}, & \tau^{(k)} &= \frac{1}{4(B_X + \rho)} \epsilon^{(k)}, \end{aligned} \quad (2.46)$$

where $\max\{\|X_*\|_F, \|X^{(k)}\|_F\} \leq B_X$ for all $k \geq 1$. For this specific choice of $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k \in \mathbb{Z}_+}$ sequence, we have $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} = \frac{\epsilon^{(1)}}{(\lambda^{(1)})^2}$ for all $k \geq 1$. Hence, setting $B = \frac{\epsilon^{(1)}}{(\lambda^{(1)})^2}$, Theorem 2.4 guarantees there exist $c_2 > 0$ and $c_3 > 0$ such that for all $k \geq 1$,

$$\begin{aligned} &\left| (\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta) - (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta) \right| \\ &\leq \left(c_2 \lambda^{(1)} + c_3 \sqrt{\epsilon^{(1)}} \right) \nu^{k-1}. \end{aligned} \quad (2.47)$$

Thus,

$$\left| (\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta) - (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta) \right| \leq \epsilon,$$

for all $k \in \mathbb{Z}_+$ such that

$$k > \frac{\ln \left(\frac{c_2 \lambda^{(1)} + c_3 \sqrt{\epsilon^{(1)}}}{\epsilon} \right)}{\ln \left(\frac{1}{\nu} \right)} + 1. \quad (2.48)$$

Moreover, Theorem 2.4 also implies that there exists $c_1 > 0$ such that for $k \geq 1$,

$$\|\mathcal{A}(X^{(k)}) + y^{(k)} - b\|_2 \leq c_1 \lambda^{(1)} \nu^{k-1}.$$

Thus, $\|\mathcal{A}(X^{(k)}) + y^{(k)} - b\|_2 \leq \epsilon$ for all $k \in \mathbb{Z}_+$ such that

$$k > \frac{\ln\left(\frac{c_1 \lambda^{(1)}}{\epsilon}\right)}{\ln\left(\frac{1}{\nu}\right)} + 1. \quad (2.49)$$

Let N_{FALC} denote the number of FALC iterations required to compute an ϵ -feasible and ϵ -optimal solution. Let

$$U = \max\left\{c_2 \lambda^{(1)} + c_3 \sqrt{\epsilon^{(1)}}, c_1 \lambda^{(1)}\right\}.$$

Then (2.48) and (2.49) imply that for all $\epsilon > 0$,

$$N_{\text{FALC}} \leq \frac{\ln\left(\frac{U}{\epsilon}\right)}{\ln\left(\frac{1}{\nu}\right)} + 1. \quad (2.50)$$

and from Lemma 2.3 it follows that N_{FALC} FALC iterations require at most

$$N_{op} = \mathcal{O}\left(\sum_{k=1}^{N_{\text{FALC}}} \frac{\min\{nm^2, n^2m\}}{\sqrt{\epsilon^{(k)}}}\right)$$

operations. Since

$$\sum_{k=1}^{N_{\text{FALC}}} \sqrt{\frac{1}{\epsilon^{(k)}}} = \frac{1}{\sqrt{\epsilon^{(1)}}} \sum_{k=1}^{N_{\text{FALC}}} \nu^{-k} = \frac{\nu^{-N_{\text{FALC}}}}{(1-\nu)} \cdot \frac{1}{\sqrt{\epsilon^{(1)}}},$$

it follows that

$$N_{op} = \mathcal{O}\left(\frac{\min\{nm^2, n^2m\}}{\sqrt{\epsilon^{(1)}}} \cdot \nu^{-N_{\text{FALC}}}\right).$$

From (2.50) it follows that for all $\epsilon > 0$ an ϵ -feasible and ϵ -optimal solution can be computed in at most

$$N_{op} = \mathcal{O}\left(\frac{U \min\{nm^2, n^2m\}}{\nu \sqrt{\epsilon^{(1)}}} \cdot \frac{1}{\epsilon}\right),$$

operations. \square

3. Implementation Details of Algorithm FALC. In this section we describe all the details of FALC. The implementable version of FALC for solving problem (1.1) with $\|\sigma(\cdot)\|_\alpha$ and $\|\cdot\|_\beta$ denoting the nuclear and ℓ_1 norms, respectively, is shown in Figure 3.1.

3.1. Bounds on the iterates $\{X^{(k)}\}_{k \in \mathbb{Z}_+}$. Let $X^{(0)} = \operatorname{argmin}_{X \in \mathbb{R}^{m \times n}} \{\|X\|_F : \mathcal{A}(X) = b\}$, $s^{(0)} = d - \mathcal{C}(X^{(0)})$ and $y^{(0)} = 0$, where \mathcal{A} is a surjective linear map. Computing $X^{(0)}$ requires a projection onto the affine space $\{X \in \mathbb{R}^{m \times n} : \mathcal{A}(X) = b\}$. Let $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) = \operatorname{argmin}\{P^{(k)}(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$. Since $\mathcal{A}(X^{(0)}) = b$, $s^{(0)} = d - \mathcal{C}(X^{(0)})$ and $f^{(k)}(X, s, y) \geq 0$ for all $X \in \mathbb{R}^{m \times n}$, $s \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, it follows that

$$\mu_1 \|\sigma(X_*^{(k)})\|_\alpha + \mu_2 \|s_*^{(k)}\|_\beta \leq \mu_1 \|\sigma(X^{(0)})\|_\alpha + \mu_2 \|s^{(0)}\|_\beta + \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2 \right). \quad (3.1)$$

Let $\eta := \mu_1 \|\sigma(X^{(0)})\|_\alpha + \mu_2 \|s^{(0)}\|_\beta$ and for each $k \geq 1$, $\eta^{(k)} := \frac{\lambda^{(k)}}{2} \left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2 \right)$. We inexactly minimize $P^{(k)}$ over the set

$$\left\{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, y \in \mathbb{R}^q : \mu_1 \|\sigma(X)\|_\alpha \leq \eta^{(k)}, \|y\|_\gamma \leq \rho\right\}, \quad (3.2)$$

to ensure that the $\{X^{(k)}\}_{k \in \mathbb{Z}_+}$ sequence always remains bounded, which also implies that $\{s^{(k)}\}_{k \in \mathbb{Z}_+}$ sequence also remains bounded (see (2.45)).

FIRST-ORDER AUGMENTED LAGRANGIAN ALGORITHM $(\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k \in \mathbb{Z}_+})$

$L \leftarrow \sigma_{max}^2(M), \quad X^{(0)} \leftarrow \operatorname{argmin}\{\|X\|_F \mid \mathcal{A}(X) = b\}, \quad s^{(0)} \leftarrow d - \mathcal{C}(X^{(0)}), \quad y^{(0)} \leftarrow 0, \quad \tau^{(0)} \leftarrow \infty, \quad k \leftarrow 0$

while $(k \geq 0)$

do

$k \leftarrow k + 1$

 INITIALIZE()

$\ell \leftarrow 0, \quad \Sigma_1 \leftarrow 0, \quad \Sigma_2 \leftarrow 0, \quad \Sigma_3 \leftarrow 0, \quad \vartheta^{(0)} \leftarrow 1$

repeat

$(X_3^{(k,\ell)}, s_3^{(k,\ell)}, y_3^{(k,\ell)}) \leftarrow (1 - \vartheta^{(\ell)})(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)}) + \vartheta^{(\ell)}(X_2^{(k,\ell)}, s_2^{(k,\ell)}, y_2^{(k,\ell)})$

$\Sigma_1 \leftarrow \Sigma_1 + \frac{\nabla_X f^{(k)}(X_3^{(k,\ell)}, s_3^{(k,\ell)}, y_3^{(k,\ell)})}{\vartheta^{(\ell)}}$

$\Sigma_2 \leftarrow \Sigma_2 + \frac{\nabla_s f^{(k)}(X_3^{(k,\ell)}, s_3^{(k,\ell)}, y_3^{(k,\ell)})}{\vartheta^{(\ell)}}$

$\Sigma_3 \leftarrow \Sigma_3 + \frac{\nabla_y f^{(k)}(X_3^{(k,\ell)}, s_3^{(k,\ell)}, y_3^{(k,\ell)})}{\vartheta^{(\ell)}}$

$[U, D, V] = \operatorname{svd}(X^{(k,0)} - \frac{\Sigma_1}{L_p}),$

$d = \operatorname{diag}(D), \quad d_\pi = \Pi_{\ell_1} \left(d, \frac{\lambda^{(k)} \mu_1}{L_p (\vartheta^{(\ell)})^2}, \frac{\eta^{(k)}}{\mu_1} \right), \quad d_+ = \left(d - \frac{\lambda^{(k)} \mu_1}{L_p (\vartheta^{(\ell)})^2} \right)^+$

1 $X_2^{(k,\ell+1)} \leftarrow U \operatorname{diag}(d_\pi) V^T, \quad \check{X}_2^{(k,\ell+1)} \leftarrow U \operatorname{diag}(d_+) V^T$

2 $s_2^{(k,\ell+1)} \leftarrow \operatorname{sign} \left(s^{(k,0)} - \frac{\Sigma_2}{L_p} \right) \odot \left(s^{(k,0)} - \frac{\Sigma_2}{L_p} - \frac{\lambda^{(k)} \mu_2}{L_p (\vartheta^{(\ell)})^2} \right)^+$

3 $y_2^{(k,\ell+1)} \leftarrow \operatorname{argmin} \left\{ \|y - \left(s^{(k,0)} - \frac{\Sigma_2}{L_p} \right)\|_2 : \|y\|_\gamma \leq \rho \right\}$

4 $(X_1^{(k,\ell+1)}, s_1^{(k,\ell+1)}, y_1^{(k,\ell+1)}) \leftarrow (1 - \vartheta^{(\ell)})(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)}) + \vartheta^{(\ell)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})$

5 $G \leftarrow (\vartheta^{(\ell)})^2 \left(L_p (X_1^{(k,0)} - \check{X}_2^{(k,\ell+1)}) - \Sigma_1 \right) + \nabla f_X^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})$

6 $g \leftarrow \operatorname{argmin}\{\|v\|_2 : v = \lambda^{(k)} \mu_2 p + \nabla_s f^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)}), p \in \partial \|\cdot\|_1|_{s_2^{(k,\ell+1)}}\}$

7 $\phi \leftarrow \rho \|\nabla_y f^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})\|_{\gamma^*} + \nabla_y f^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})^T y_2^{(k,\ell+1)}$

$\vartheta^{(\ell+1)} \leftarrow \frac{\sqrt{(\vartheta^{(\ell)})^4 - 4(\vartheta^{(\ell)})^2} - (\vartheta^{(\ell)})^2}{2}$

if $\|X_1^{(k,\ell+1)} - X_1^{(k,\ell)}\|_F \leq \varrho$ **and** $\|s_1^{(k,\ell+1)} - s_1^{(k,\ell)}\|_2 \leq \varrho$

then

$(X_{sol}, s_{sol}) \leftarrow (X_1^{(k,\ell+1)}, s_1^{(k,\ell+1)})$

return (X_{sol}, s_{sol})

if $(\ell == 0)$

then

$\tau_X^{(k)} \leftarrow \min\{\bar{c}_\tau \tau_X^{(k-1)}, c_\tau \|G\|_F\}, \quad \tau_s^{(k)} \leftarrow \min\{\bar{c}_\tau \tau_s^{(k-1)}, c_\tau \|g\|_2\}$

$\xi^{(k)} \leftarrow \min\{\bar{c}_\xi \xi^{(k-1)}, c_\xi \phi\}$

$\ell \leftarrow \ell + 1$

until $(\ell > N^{(k)})$ **or** $(\|G\|_F \leq \tau_X^{(k)} \text{ and } \|g\|_2 \leq \tau_s^{(k)} \text{ and } \phi \leq \xi^{(k)})$

if $(\|G\|_F \leq \tau_X^{(k)} \text{ and } \|g\|_2 \leq \tau_s^{(k)} \text{ and } \phi \leq \xi^{(k)})$

then

$(X^{(k)}, s^{(k)}, y^{(k)}) \leftarrow (X_2^{(k,\ell)}, s_2^{(k,\ell)}, y_2^{(k,\ell)})$

else

$(X^{(k)}, s^{(k)}, y^{(k)}) \leftarrow (X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})$

$\theta_1^{(k+1)} \leftarrow \theta_1^{(k)} - \frac{\mathcal{A}(X^{(k)}) + y^{(k)} - b}{\lambda^{(k)}}$

$\theta_2^{(k+1)} \leftarrow \theta_2^{(k)} - \frac{\mathcal{C}(X^{(k)}) + s^{(k)} - d}{\lambda^{(k)}}$

FIG. 3.1. Details of the First-Order Augmented Lagrangian Algorithm (FALC)

INITIALIZE ()

$$\begin{aligned}
f^{(k)}(X, s, y) &= \frac{1}{2}\|\mathcal{A}(X) + y - b - \lambda^{(k)}\theta_1^{(k)}\|_2^2 + \frac{1}{2}\|\mathcal{C}(X) + s - d - \lambda^{(k)}\theta_2^{(k)}\|_2^2 \\
\beta^{(k)} &\leftarrow \mu_1\|X^{(0)}\|_* + \mu_2\|s^{(0)}\|_1 + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right) \\
N^{(k)} &\leftarrow \left\lfloor \sigma_{\max}(M) \sqrt{\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2}} \left(\beta^{(k)} + \mu_1\|X^{(k-1)}\|_* + \mu_2\|s^{(k-1)}\|_1 \right) \sqrt{\frac{2}{\epsilon^{(k)}}} \right\rfloor \\
X_1^{(k,0)} &\leftarrow X^{(k-1)}, \quad X_2^{(k,0)} \leftarrow X^{(k-1)}, \quad X_3^{(k,0)} \leftarrow X^{(k-1)} \\
s_1^{(k,0)} &\leftarrow s^{(k-1)}, \quad s_2^{(k,0)} \leftarrow s^{(k-1)}, \quad s_3^{(k,0)} \leftarrow s^{(k-1)} \\
y_1^{(k,0)} &\leftarrow y^{(k-1)}, \quad y_2^{(k,0)} \leftarrow y^{(k-1)}, \quad y_3^{(k,0)} \leftarrow y^{(k-1)}
\end{aligned}$$

FIG. 3.2. Details of INITIALIZE subroutine

3.2. Subgradient selection. In order to check the stopping condition (2.3)(b), in line 5 and line 6 of Figure 3.1 we compute a subgradient $(G, g) \in \partial_{X,s} P^{(k)}(\cdot, \cdot, \cdot)|_{(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})}$ such that $G = \lambda^{(k)}\mu_1 Q + \nabla_X f^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})$ and $g = \lambda^{(k)}\mu_2 q + \nabla_s f^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})$, where

$$\begin{aligned}
Q &= \frac{(\vartheta^{(\ell)})^2 L_p}{\lambda^{(k)}\mu_1} \left(X_1^{(k,0)} - \check{X}_2^{(k,\ell+1)} - \frac{1}{L_p} \sum_{i=0}^{\ell} \frac{\nabla_X f^{(k)}(X_3^{(k,i)}, s_3^{(k,i)}, y_3^{(k,i)})}{\vartheta^{(i)}} \right), \\
q &= \operatorname{argmin} \left\{ \|\lambda^{(k)}\mu_2 r + \nabla_s f^{(k)}(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})\|_2 : r \in \partial\|\cdot\|_{\beta}|_{s_2^{(k,\ell+1)}} \right\},
\end{aligned}$$

where $\check{X}_2^{(k,\ell+1)}$ is defined in (3.6). It can be easily shown that $Q \in \partial\|\sigma(\cdot)\|_{\alpha}|_{X_2^{(k,\ell+1)}}$, and, given $\nabla_s f^{(k)}$ at $(X_2^{(k,\ell+1)}, s_2^{(k,\ell+1)}, y_2^{(k,\ell+1)})$ the complexity of computing $q \in \partial\|\cdot\|_{\beta}|_{s_2^{(k,\ell+1)}} \subset \mathbb{R}^p$ is $\mathcal{O}(p)$ when $\beta \in \{1, 2\}$ and $\mathcal{O}(p \log(p))$ when $\beta = \infty$.

3.3. Details of inner iterations. We now discuss the details of Algorithm 3 [33] for solving

$$\min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_{\gamma} \leq \rho} \lambda(\mu_1\|\sigma(X)\|_{\alpha} + \mu_2\|s\|_{\beta}) + f(X, s, y), \quad (3.3)$$

where f is a proper, lower semicontinuous (lsc), convex function and has a Lipschitz continuous gradient ∇f with constant L with respect to norm $\|\cdot\|$ on $\mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q$ such that for any $X \in \mathbb{R}^{m \times n}$, $s \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$,

$$\|(X, s, y)\| = \sqrt{\|X\|_F^2 + \|s\|_2^2 + \|y\|_2^2}. \quad (3.4)$$

Let $(X_*, s_*, y_*) = \operatorname{argmin}_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_{\gamma} \leq \rho} \lambda(\mu_1\|\sigma(X)\|_{\alpha} + \mu_2\|s\|_{\beta}) + f(X, s, y)$ and let η denote any upper bound $\|\sigma(X_*)\|_{\alpha}$ of the unconstrained optimal solution X_* . Algorithm 3 computes three sets of iterates for all the variables $(X_1^{(\ell)}, X_2^{(\ell)}, X_3^{(\ell)})$, $(s_1^{(\ell)}, s_2^{(\ell)}, s_3^{(\ell)})$ and $(y_1^{(\ell)}, y_2^{(\ell)}, y_3^{(\ell)})$:

1. $(X_3^{(\ell)}, s_3^{(\ell)}, y_3^{(\ell)})$ is a convex combination of $(X_1^{(\ell)}, s_1^{(\ell)}, y_1^{(\ell)})$ and $(X_2^{(\ell)}, s_2^{(\ell)}, y_2^{(\ell)})$:

$$\begin{aligned}
X_3^{(\ell)} &= (1 - \vartheta^{(\ell)})X_1^{(\ell)} + \vartheta^{(\ell)}X_2^{(\ell)}, \\
s_3^{(\ell)} &= (1 - \vartheta^{(\ell)})s_1^{(\ell)} + \vartheta^{(\ell)}s_2^{(\ell)}, \\
y_3^{(\ell)} &= (1 - \vartheta^{(\ell)})y_1^{(\ell)} + \vartheta^{(\ell)}y_2^{(\ell)}.
\end{aligned}$$

2. $(X_2^{(\ell+1)}, s_2^{(\ell+1)}, y_2^{(\ell+1)})$ is computed using the gradients $\nabla f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)})$ for all the iterates $i \leq \ell$:

$$\begin{aligned} X_2^{(\ell+1)} &= \operatorname{argmin}_{X: \|\sigma(X)\|_\alpha \leq \eta} \left\{ \sum_{i=0}^{\ell} \frac{\langle \nabla_X f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)}), X \rangle + \lambda \mu_1 \|\sigma(X)\|_\alpha}{\vartheta^{(i)}} + \frac{L}{2} \|X - X_1^{(0)}\|_F^2 \right\}, \\ &= \operatorname{argmin}_{X: \|\sigma(X)\|_\alpha \leq \eta} \left\{ \frac{1}{2} \left\| X - \left(X_1^{(0)} - \frac{1}{L} \sum_{i=0}^{\ell} \frac{\nabla_X f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)})}{\vartheta^{(i)}} \right) \right\|_F^2 \right. \\ &\quad \left. + \left(\sum_{i=0}^{\ell} \frac{\lambda \mu_1}{\vartheta^{(i)} L} \right) \|\sigma(X)\|_\alpha \right\}. \end{aligned} \quad (3.5)$$

$\check{X}_2^{(\ell+1)}$ is the unconstrained solution of the problem (3.5), i.e.

$$\begin{aligned} \check{X}_2^{(\ell+1)} &= \operatorname{argmin}_X \left\{ \frac{1}{2} \left\| X - \left(X_1^{(0)} - \frac{1}{L} \sum_{i=0}^{\ell} \frac{\nabla_X f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)})}{\vartheta^{(i)}} \right) \right\|_F^2 \right. \\ &\quad \left. + \left(\sum_{i=0}^{\ell} \frac{\lambda \mu_1}{\vartheta^{(i)} L} \right) \|\sigma(X)\|_\alpha \right\}, \end{aligned} \quad (3.6)$$

and $\check{X}_2^{(\ell+1)}$ is used for computing subgradients as explained in Section 3.2.

Similarly,

$$\begin{aligned} s_2^{(\ell+1)} &= \operatorname{argmin}_s \left\{ \sum_{i=0}^{\ell} \frac{\langle \nabla_s f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)}), s \rangle + \lambda \mu_2 \|s\|_\beta}{\vartheta^{(i)}} + \frac{L}{2} \|s - s_1^{(0)}\|_2^2 \right\}, \\ &= \operatorname{argmin}_s \left\{ \frac{1}{2} \left\| s - \left(s_1^{(0)} - \frac{1}{L} \sum_{i=0}^{\ell} \frac{\nabla_s f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)})}{\vartheta^{(i)}} \right) \right\|_2^2 + \left(\sum_{i=0}^{\ell} \frac{\lambda \mu_2}{\vartheta^{(i)} L} \right) \|s\|_\beta \right\}. \end{aligned} \quad (3.7)$$

and

$$\begin{aligned} y_2^{(\ell+1)} &= \operatorname{argmin}_{y: \|y\|_\gamma \leq \rho} \left\{ \sum_{i=0}^{\ell} \frac{\langle \nabla_y f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)}), y \rangle}{\vartheta^{(i)}} + \frac{L}{2} \|y - y_1^{(0)}\|_2^2 \right\}, \\ &= \operatorname{argmin}_{y: \|y\|_\gamma \leq \rho} \left\{ \frac{1}{2} \left\| y - \left(y_1^{(0)} - \frac{1}{L} \sum_{i=0}^{\ell} \frac{\nabla_y f(X_3^{(i)}, s_3^{(i)}, y_3^{(i)})}{\vartheta^{(i)}} \right) \right\|_2^2 \right\}. \end{aligned} \quad (3.8)$$

Thus, in Step 1 and Step 2 of Figure 3.1 we compute the $X_2^{(\ell)}$ and $s_2^{(\ell)}$ iterates, respectively.

3. $(X_1^{(\ell+1)}, s_1^{(\ell+1)}, y_1^{(\ell+1)})$ is a convex combination of $(X_1^{(\ell)}, s_1^{(\ell)}, y_1^{(\ell)})$ and $(X_2^{(\ell+1)}, s_2^{(\ell+1)}, y_2^{(\ell+1)})$:

$$\begin{aligned} X_1^{(\ell+1)} &= (1 - \vartheta^{(\ell)}) X_1^{(\ell)} + \vartheta^{(\ell)} X_2^{(\ell+1)}, \\ s_1^{(\ell+1)} &= (1 - \vartheta^{(\ell)}) s_1^{(\ell)} + \vartheta^{(\ell)} s_2^{(\ell+1)}, \\ y_1^{(\ell+1)} &= (1 - \vartheta^{(\ell)}) y_1^{(\ell)} + \vartheta^{(\ell)} y_2^{(\ell+1)}. \end{aligned}$$

In order to solve (3.5) and (3.6) we need to compute the SVD of an appropriately defined matrix. The iteration description above implicitly assumed that we need to compute this SVD exactly. This is not necessary – inexactly computing the SVD adds a small additional error term.

3.4. Stopping criterion for FALC. In our numerical experiments, we terminate either the distance between successive inner iterates are below a threshold ϱ for each component, i.e. $\|X_1^{(k,\ell)} - X_1^{(k,\ell-1)}\|_F \leq \varrho$, $\|s_1^{(k,\ell)} - s_1^{(k,\ell-1)}\|_2 \leq \varrho$ or there exist partial subgradients with sufficiently small norm for each component, i.e. $\|G\|_F \leq \varsigma_X$, $\|g\|_2 \leq \varsigma_s$ for some $(G, g) \in \partial_{X,s} P^{(k)}(\cdot, \cdot, \cdot)|_{(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})}$ and

$$\rho \|\nabla_y P^{(k)}(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})\|_{\gamma^*} + \nabla_y P^{(k)}(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})^T y^{(k)} \leq \varsigma_y.$$

In our numerical experiments we set ϱ , ς_X , ς_s and ς_y by experimenting with a small instance of the problem.

3.5. Multiplier selection. Given $\bar{c}_\tau \in (0, 1)$, $\bar{c}_\xi \in (0, 1)$, $\bar{c}_\lambda > 0$, $c_\tau \in (0, 1)$, $c_\xi \in (0, 1)$, $c_\lambda \in (0, 1)$, for all $k \geq 1$ the approximate optimality parameters $\tau^{(k)}$, $\xi^{(k)}$ and the penalty parameter $\lambda^{(k)}$ are set as follows:

$$\begin{aligned}
Z^{(1)} &= \operatorname{argmin}_{Z \in \mathbb{R}^{m \times n}} \|Z - \left(X^{(0)} - \frac{1}{L_p} \nabla_X f^{(1)}(X^{(0)}, s^{(0)}, y^{(0)})\right)\|_F^2 + \frac{\lambda^{(1)} \mu_1}{L_p} \|\sigma(Z)\|_\alpha, \\
z^{(1)} &= \operatorname{argmin}_{z \in \mathbb{R}^p} \|z - \left(s^{(0)} - \frac{1}{L_p} \nabla_s f^{(1)}(X^{(0)}, s^{(0)}, y^{(0)})\right)\|_2^2 + \frac{\lambda^{(1)} \mu_2}{L_p} \|z\|_\beta, \\
v^{(1)} &= \operatorname{argmin}_{v: \|v\|_\gamma \leq \rho} \|v - \left(y^{(0)} - \frac{1}{L_p} \nabla_y f^{(1)}(X^{(0)}, s^{(0)}, y^{(0)})\right)\|_2^2, \\
G^{(1)} &= L_p \left(X^{(0)} - \frac{1}{L_p} \nabla_X f^{(1)}(X^{(0)}, s^{(0)}, y^{(0)}) - Z^{(1)}\right) + \nabla_X f^{(1)}(Z^{(1)}, z^{(1)}, v^{(1)}), \\
g^{(1)} &= \operatorname{argmin}\{\|v\|_2 : v = \lambda^{(1)} \mu_2 p + \nabla_s f^{(1)}(Z^{(1)}, z^{(1)}, v^{(1)}), p \in \partial \|\cdot\|_\beta|_{z^{(1)}}\}, \\
\tau_X^{(1)} &= \bar{c}_\tau \|G^{(1)}\|_F, \\
\tau_s^{(1)} &= \bar{c}_\tau \|g^{(1)}\|_2, \\
\xi^{(1)} &= \bar{c}_\xi (\rho \|\nabla_y f^{(1)}(Z^{(1)}, z^{(1)}, v^{(1)})\|_{\gamma^*} + \nabla_y f^{(1)}(Z^{(1)}, z^{(1)}, v^{(1)})^T v^{(1)}) \\
\lambda^{(1)} &= \bar{c}_\lambda \|X^{(0)}\|_2, \\
Z^{(k)} &= \operatorname{argmin}_{Z \in \mathbb{R}^{m \times n}} \|Z - \left(X^{(k-1)} - \frac{1}{L_p} \nabla_X f^{(k)}(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})\right)\|_F^2 + \frac{\lambda^{(k)} \mu_1}{L_p} \|\sigma(Z)\|_\alpha, \\
z^{(k)} &= \operatorname{argmin}_{z \in \mathbb{R}^p} \|z - \left(s^{(k-1)} - \frac{1}{L_p} \nabla_s f^{(k)}(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})\right)\|_2^2 + \frac{\lambda^{(k)} \mu_2}{L_p} \|z\|_\beta, \\
v^{(k)} &= \operatorname{argmin}_{v: \|v\|_\gamma \leq \rho} \|v - \left(y^{(k-1)} - \frac{1}{L_p} \nabla_y f^{(k)}(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})\right)\|_2^2, \\
G^{(k)} &= L_p \left(X^{(k-1)} - \frac{1}{L_p} \nabla_X f^{(k)}(X^{(k-1)}, s^{(k-1)}, y^{(k-1)}) - Z^{(k)}\right) + \nabla_X f^{(k)}(Z^{(k)}, z^{(k)}, v^{(k)}), \\
g^{(k)} &= \operatorname{argmin}\{\|v\|_2 : v = \lambda^{(k)} \mu_2 p + \nabla_s f^{(k)}(Z^{(k)}, z^{(k)}, v^{(k)}), p \in \partial \|\cdot\|_\beta|_{z^{(k)}}\}, \\
\tau_X^{(k)} &= \min\{c_\tau \tau_X^{(k-1)}, \bar{c}_\tau \|G^{(k)}\|_F\}, \\
\tau_s^{(k)} &= \min\{c_\tau \tau_s^{(k-1)}, \bar{c}_\tau \|g^{(k)}\|_2\}, \\
\xi^{(k)} &= \min\{c_\xi \xi^{(k-1)}, (\rho \|\nabla_y f^{(k)}(Z^{(k)}, z^{(k)}, v^{(k)})\|_{\gamma^*} + \nabla_y f^{(k)}(Z^{(k)}, z^{(k)}, v^{(k)})^T v^{(k)})\} \\
\lambda^{(k)} &= c_\lambda \lambda^{(k-1)},
\end{aligned} \tag{3.9}$$

for all $k \geq 2$. In all our experiments, $\bar{c}_\tau = 0.999$ and $\bar{c}_\xi = 0.999$.

We initialize FALC with $(X^{(0)}, s^{(0)})$ such that $\mathcal{A}(X^{(0)}) = b$ and $s^{(0)} = d - \mathcal{C}(X^{(0)})$. In first iteration of FALC, we solve the problem

$$\min_{\|\sigma(X)\|_\alpha \leq \eta^{(1)}} P^{(1)}(X, s) = \min_{\|\sigma(X)\|_\alpha \leq \eta^{(1)}} \lambda^{(1)} (\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f^{(1)}(X, s),$$

where $\eta^{(1)} = \mu_1 \|\sigma(X^{(0)})\|_\alpha + \mu_2 \|s^{(0)}\|_\beta$. Since $X^{(0)}$ is feasible, $f^{(1)}(X^{(0)}, s^{(0)}) = 0$ and $P^{(1)}(X^{(0)}, s^{(0)}) = \lambda^{(1)} \eta^{(1)}$ and $P^{(1)}(X) \geq 0$ for all $X \in \mathbb{R}^{m \times n}$, the initial duality gap is less than or equal to $\lambda^{(1)} \beta^{(1)}$. Hence, we initialize $\epsilon^{(1)} = 0.99 \lambda^{(1)} \eta^{(1)}$ and then set $\epsilon^{(k+1)} = c_\lambda^2 \epsilon^{(k)}$ for all $k \geq 1$.

4. Numerical experiments. We conducted two sets of numerical experiments with FALC. In the first set of experiments we solved a set of randomly generated instances of principle component pursuit problems (1.5). In this setting, we compare FALC with other augmented Lagrangian algorithms I-ALM [25], APG [26] and soft-thresholding algorithm SVT [4]. In the second set of experiments, we solved a set of randomly generated instances of stable principle component pursuit problem (1.6) using only FALC. In Section 4.1, we describe the methodology we have used in both experimental settings for generating random problem instances.

4.1. Data generation. We tested FALC on randomly generated data matrices $D = X_0 + S_0 + Y_0$, where

- i. $X_0 = UV^T$, such that $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{n \times r}$ for $r = 0.05n$ and $U_{ij} \sim \mathcal{N}(0, 1)$, $V_{ij} \sim \mathcal{N}(0, 1)$ for all i, j are independent standard Gaussian variables,
- ii. $\Lambda \subset \{(i, j) : 1 \leq i, j \leq n\}$ such that cardinality of Λ , $|\Lambda| = p$ for $p = 0.05n^2$,
- iii. $(S_0)_{ij} \sim \mathcal{U}[-1, 1]$ for all $(i, j) \in \Lambda$ are independent uniform random variables between -1 and 1 ,
- iv. $(Y_0)_{ij} \sim \rho \mathcal{U}[-1, 1]$ for all i, j are independent Gaussian variables.

4.2. Principle Component Pursuit Problem. In this section we solve the problem

$$\begin{aligned}
&\min_{X, S \in \mathbb{R}^{m \times n}} \|X\|_* + \mu_2 \|\mathbf{vec}(S)\|_1, \\
&\text{subject to } X + S = D,
\end{aligned} \tag{4.1}$$

and report the results of our numerical experiments comparing FALC with I-ALM [25], APG [26] and SVT [4]. All the codes for I-ALM, APG and SVT, can be found at <http://perception.csl.uiuc.edu/matrix-rank/home.html>. Please note that SVT [4] algorithm was originally proposed for solving the matrix completion problem. The algorithm we used in our numerical study is an adaptation of the SVT algorithm by Wright and Rao at the Perception and Decision Laboratory in University of Illinois, Urbana-Champaign to solve robust PCA problem.

We created 10 random problems of size $n = 500$, i.e. $D \in \mathbb{R}^{500 \times 500}$ using the procedure described in Section 4.1, where ρ is set to 0, i.e. $Y_0 = 0$. We chose parameter values for each of the four algorithms so that they produce a solution X_{sol} and S_{sol} with relative-infeasibility approximately equal to 5×10^{-9} , i.e. $\frac{\|X_{sol} + S_{sol} - D\|_F}{\|D\|_F} \approx 5 \times 10^{-9}$. For each algorithm we set the parameters by solving a set of small size problems and these parameter values were fixed throughout the experiments, all other parameters are set to their default values. The termination criteria are not directly comparable due to different formulations of the problem solved by different solvers. For FALC we attempted to set the stopping parameter ϱ such that on average the stopping criterion for FALC is tighter than the stopping criteria of all the other algorithms we tested.

1. **FALC**: Problem (4.1) is a special case of Problem (1.1) with $\rho = 0$. Therefore, $f^{(k)}(X, s, y)$ defined in (2.2) simplifies to $f^{(k)}(X, S) = \frac{1}{2} \|\mathbf{vec}(X + S) - \mathbf{vec}(D) - \lambda^{(k)} \theta_1^{(k)}\|_2^2$ (note that for all $k \geq 1$, $\theta_2^{(k)} = 0$). We set $c_\tau = 0.4$, $c_\xi = 0.4$, $c_\lambda = 0.4$, $\bar{c}_\tau = 0.999$, $\bar{c}_\xi = 0.999$, $\bar{c}_\lambda = 2$ and initialize $\theta_1^{(1)}$ as in [25], i.e.

$$\theta_1^{(1)} = \frac{1}{\max\{\|sign(D)\|_2, \sqrt{n} \|\mathbf{vec}(sign(D))\|_\infty\}} \mathbf{vec}(sign(D)). \quad (4.2)$$

Finally, we set $\varrho = 1 \times 10^{-5}$ and terminate FALC when the distance between successive inner iterates are below the threshold ϱ for each component, i.e. $\|X_1^{(k, \ell)} - X_1^{(k, \ell-1)}\|_F \leq \varrho$ and $\|s_1^{(k, \ell)} - s_1^{(k, \ell-1)}\|_2 \leq \varrho$ for any $k \geq 1$. We used PROPACK [23] for computing partial singular value decompositions. In order to estimate the rank of X_0 , we followed the scheme proposed in Equation (17) in [25]. The code for PROPACK is available at [<http://soi.stanford.edu/~rmunk/PROPACK/>].

2. **I-ALM**: I-ALM solves $\min\{\|X\|_* + \frac{1}{\sqrt{n}} \|\mathbf{vec}(S)\|_1 : X + S = D\}$. Let $(X^{(k)}, s^{(k)})$ be the k -th iterate. I-ALM terminates when $\frac{\|X^{(k)} + s^{(k)} - D\|_F}{\|D\|_F} \leq 1 \times 10^{-8}$.
3. **APG**: For some $\bar{\lambda} > 0$, APG solves $\min\left\{\bar{\lambda} \left(\|X\|_* + \frac{1}{\sqrt{n}} \|\mathbf{vec}(S)\|_1\right) + \frac{1}{2} \|X + S - D\|_F^2\right\}$. Stopping tolerance is set to 5×10^{-11} (the definition of stopping criteria is complicated, for details see partial APG code at [<http://perception.csl.uiuc.edu/matrix-rank/home.html>]). In the code, by default $\bar{\lambda}$ is set to $\sigma_{\max}(D) \times 10^{-9}$.
4. **SVT**: SVT solves a relaxation of the robust PCA problem, $\min\left\{\bar{\lambda} \left(\|X\|_* + \frac{1}{\sqrt{n}} \|\mathbf{vec}(S)\|_1\right) + \frac{1}{2} (\|X\|_F^2 + \|S\|_F^2) : X + S = D\right\}$. Let $(X^{(k)}, s^{(k)})$ be the k -th iterate when $\bar{\lambda}$ is set to 1×10^3 . SVT terminates $\frac{\|X^{(k)} + s^{(k)} - D\|_F}{\|D\|_F} \leq 5 \times 10^{-4}$. Please note that we have chosen a weaker stopping criterion for SVT.

The results of the experiments are displayed in Tables 4.1–4.2. In Table 4.1–4.2, the row labeled **CPU** lists the running time of each algorithm in *seconds* and all other rows are self-explanatory. The column labeled **average** lists the average taken over the $N = 10$ random instances, the columns labeled **min** (resp. **max**) list the minimum (resp. maximum) over the 10 instances. The experimental results in Table 4.1–4.2, show that FALC is competitive with the state of the art algorithms, e.g. I-ALM, APG and SVT, specialized for solving robust PCA problem. Even though FALC is not special purpose algorithm for robust PCA, in our numerical experiments, we requires FALC requires fewer singular value decompositions when compared to APG and SVT. In addition, for all 10 randomly created problems in the test set, only FALC and I-ALM accurately identified the zero-set of the sparse component S_0 , i.e. $I_0 = \{(ij) \in \{1, 2, \dots, n\} \times \{1, 2, \dots, n\} : (S_0)_{ij} = 0\}$ without any thresholding. This feature of FALC is very appealing in practice. For signals with a large dynamic range, almost all of the state-of-the-art efficient algorithms produce a solution with many small non zeros terms, and it is often hard to determine the threshold.

TABLE 4.1
Experiment Results for $n = 500$, $r = 0.05n^2$, $p = 0.05n$ and $\frac{\|X_{sol} + S_{sol} - D\|_F}{\|D\|_F} \approx 5 \times 10^{-9}$

	FALC			I-ALM		
	Average	Min	Max	Average	Min	Max
svd #	42.6	40	45	31.6	30	33
$\ X_{sol} - X_0\ _F / \ X_0\ _F$	4.65E-09	2.28E-09	7.04E-09	1.85E-09	5.90E-10	3.40E-09
$\ S_{sol} - S_0\ _F / \ S_0\ _F$	1.79E-07	8.89E-08	2.69E-07	1.94E-07	4.80E-08	3.85E-07
$\ \ X_{sol}\ _* - \ X_0\ _* \ / \ X_0\ _*$	1.88E-10	4.74E-11	4.15E-10	1.13E-11	3.67E-12	2.07E-11
$\max\{ \sigma_i - \sigma_i^0 : \sigma_i^0 > 0\}$	2.61E-07	1.01E-07	5.15E-07	8.69E-08	2.34E-08	2.54E-07
$\max\{ \sigma_i : \sigma_i^0 = 0\}$	1.57E-13	4.22E-14	3.48E-13	1.47E-13	5.92E-14	3.66E-13
$\ \ \text{vec}(S_{sol})\ _1 - \ \text{vec}(X_0)\ _1 \ / \ \text{vec}(X_0)\ _1$	1.97E-08	7.44E-09	3.02E-08	2.24E-09	4.13E-10	5.11E-09
$\max\{ (S_{sol})_{ij} - (S_0)_{ij} : (S_0)_{ij} > 0\}$	1.31E-06	5.99E-07	1.75E-06	1.07E-05	2.31E-06	2.45E-05
$\max\{ (S_{sol})_{ij} : (S_0)_{ij} = 0\}$	0	0	0	0	0	0
rank	25	25	25	25	25	25
$\ X_{sol} + S_{sol} - D\ _F / \ D\ _F$	4.67E-09	2.31E-09	7.04E-09	4.66E-09	1.08E-09	9.63E-09
CPU	23.6	19.4	32.3	15.9	12.0	24.4

TABLE 4.2
Experiment Results for $n = 500$, $r = 0.05n^2$, $p = 0.05n$ and $\frac{\|X_{sol} + S_{sol} - D\|_F}{\|D\|_F} \approx 5 \times 10^{-9}$

	APG			SVT		
	Average	Min	Max	Average	Min	Max
svd #	187.7	187	188	833.9	819	857
$\ X_{sol} - X_0\ _F / \ X_0\ _F$	4.14E-09	3.99E-09	4.39E-09	1.79E-04	1.76E-04	1.80E-04
$\ S_{sol} - S_0\ _F / \ S_0\ _F$	1.63E-07	1.57E-07	1.72E-07	2.04E-02	2.02E-02	2.08E-02
$\ \ X_{sol}\ _* - \ X_0\ _* \ / \ X_0\ _*$	3.96E-09	3.82E-09	4.20E-09	1.66E-05	1.53E-05	1.85E-05
$\max\{ \sigma_i - \sigma_i^0 : \sigma_i^0 > 0\}$	1.99E-06	1.90E-06	2.11E-06	1.45E-02	1.17E-02	1.68E-02
$\max\{ \sigma_i : \sigma_i^0 = 0\}$	1.26E-13	6.84E-14	1.91E-13	2.39E-13	7.58E-14	6.79E-13
$\ \ \text{vec}(S_{sol})\ _1 - \ \text{vec}(X_0)\ _1 \ / \ \text{vec}(X_0)\ _1$	1.83E-07	1.76E-07	1.92E-07	5.03E-03	4.89E-03	5.14E-03
$\max\{ (S_{sol})_{ij} - (S_0)_{ij} : (S_0)_{ij} > 0\}$	1.95E-07	1.80E-07	2.25E-07	1.19E-01	1.07E-01	1.33E-01
$\max\{ (S_{sol})_{ij} : (S_0)_{ij} = 0\}$	3.70E-08	2.09E-08	6.64E-08	5.50E-03	3.59E-03	8.45E-03
rank	25	25	25	25	25	25
$\ X_{sol} + S_{sol} - D\ _F / \ D\ _F$	5.43E-09	5.24E-09	5.77E-09	4.99E-04	4.98E-04	5.00E-04
CPU	87.7	71.6	101.6	265.2	252.0	273.1

4.3. Stable Principle Component Pursuit Problem.

In this section, we solve the problem

$$\begin{aligned} \min_{X, S \in \mathbb{R}^{m \times n}} \quad & \|X\|_* + \mu_2 \|\text{vec}(S)\|_1, \\ \text{subject to} \quad & \|\text{vec}(X + S - D)\|_\infty \leq \rho, \end{aligned} \quad (4.3)$$

and report the results of our numerical experiments using FALC. To best of our knowledge, there are no publicly available code specialized for solving Problem (4.3), other than general purpose SDP solvers.

We created 10 random problems of size $n = 500$, i.e. $D \in \mathbb{R}^{500 \times 500}$ using the procedure described in Section 4.1, where ρ is set to 1×10^4 , i.e. each entry of the noise term Y_0 is coming from a uniform distribution between $[-\rho, \rho]$. We chose the value of the stopping parameter so that FALC produces a solution X_{sol} and S_{sol} with $\frac{\|X_{sol} + S_{sol} - D\|_F}{\|D\|_F} \approx 1 \times 10^{-5}$.

Problem (4.3) is a special case of Problem (1.1). Therefore, $f^{(k)}(X, s, y)$ defined in (2.2) simplifies to $f^{(k)}(X, S, y) = \frac{1}{2} \|\text{vec}(X + S) + y - \text{vec}(D) - \lambda^{(k)} \theta_1^{(k)}\|_2^2$ (note that for all $k \geq 1$, $\theta_2^{(k)} = 0$). We set the parameter values for FALC by solving a set of small size problems and these parameter values were fixed throughout the experiments, all other parameters are set to their default values, i.e. $c_\tau = 0.4$, $c_\xi = 0.4$, $c_\lambda = 0.4$, $\bar{c}_\tau = 0.999$, $\bar{c}_\xi = 0.999$. We set $\bar{c}_\lambda = 1.5$ and initialize $\theta_1^{(1)}$ as in [25], i.e. as in (4.2).

Finally, We set $\varrho = 1 \times 10^{-5}$, $\varsigma = 1 \times 10^{-3}$ and terminate FALC when either the distance between successive inner iterates are below a threshold ϱ for each component, i.e. $\|\text{vec}(X_1^{(k, \ell)}) - \text{vec}(X_1^{(k, \ell-1)})\|_\infty \leq \varrho$, $\|\text{vec}(s_1^{(k, \ell)}) - \text{vec}(s_1^{(k, \ell-1)})\|_\infty \leq \varrho$ for any $k \geq 1$ or there exist partial subgradients with sufficiently small

norm for each component, i.e. $\|G\|_F \leq \varsigma/2$, $\|g\|_2 \leq \varsigma$ for some $(G, g) \in \partial_{X,s} P^{(k)}(\cdot, \cdot, \cdot)|_{(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})}$ and

$$\rho \|\nabla_y P^{(k)}(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})\|_{\gamma^*} + \nabla_y P^{(k)}(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})^T y^{(k)} \leq \varsigma.$$

We have used PROPACK [23] for computing partial singular value decompositions. In order to estimate the rank of X_0 , we followed the scheme proposed in Equation (17) in [25]. The results of the experiments are displayed in Table 4.3.

TABLE 4.3
Experiment Results for $n = 500$, $r = 0.05n^2$, $p = 0.05n$, $\rho = 1 \times 10^{-4}$ and $\frac{\|X_{\text{sol}} + S_{\text{sol}} - D\|_F}{\|D\|_F} \approx 1 \times 10^{-5}$

	FALC		
	Average	Min	Max
svd #	53.8	49	60
$\ X_{\text{sol}} - X_0\ _F / \ X_0\ _F$	1.71E-05	1.66E-05	1.83E-05
$\ S_{\text{sol}} - S_0\ _F / \ S_0\ _F$	4.04E-04	2.61E-04	9.11E-04
$\ X_{\text{sol}}\ _* - \ X_0\ _* / \ X_0\ _*$	1.59E-05	1.56E-05	1.61E-05
$\max\{ \sigma_i - \sigma_i^0 : \sigma_i^0 > 0\}$	9.70E-03	9.43E-03	1.01E-02
$\max\{ \sigma_i : \sigma_i^0 = 0\}$	1.56E-09	1.37E-10	7.97E-09
$\ \text{vec}(S_{\text{sol}})\ _1 - \ \text{vec}(X_0)\ _1 / \ \text{vec}(X_0)\ _1$	2.35E-04	2.12E-04	3.11E-04
$\max\{ \langle S_{\text{sol}} \rangle_{ij} - \langle S_0 \rangle_{ij} : \langle S_0 \rangle_{ij} > 0\}$	3.83E-03	1.26E-03	8.63E-03
$\max\{ \langle S_{\text{sol}} \rangle_{ij} : \langle S_0 \rangle_{ij} = 0\}$	0	0	0
rank	25	25	25
$\ X_{\text{sol}} + S_{\text{sol}} - D\ _F / \ D\ _F$	2.15E-05	1.95E-05	2.89E-05
CPU	35.3	30.8	44.6

5. Extensions and conclusion. The algorithmic framework proposed in this paper extends to the following much more general class of problems:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta + \langle R, X \rangle + \frac{\mu_3}{2} \|X - X_0\|_F^2, \\ \text{subject to} \quad & \mathcal{A}(X) = b, \\ & \mathcal{F}(X) \preceq G, \\ & \|\mathcal{G}(X) - h\|_\gamma \leq \rho \end{aligned} \tag{5.1}$$

where the matrix norm $\|\sigma(\cdot)\|_\alpha$ denotes either the nuclear norm, the Frobenius norm, or the ℓ_2 -norm, the vector norms $\|\cdot\|_\beta$ and $\|\cdot\|_\gamma$ denote either the ℓ_1 -norm, ℓ_2 -norm or the ℓ_∞ -norm, and $\mathcal{A}(\cdot)$, $\mathcal{C}(\cdot)$, $\mathcal{G}(\cdot)$ and $\mathcal{F}(\cdot)$ are linear operators from $\mathbb{R}^{m \times n}$ to vector spaces of appropriate dimensions. By introducing slack variables, (5.1) can be reformulated as follows.

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s_c\|_\beta + \langle R, X \rangle + \frac{\mu_3}{2} \|X - X_0\|_F^2, \\ \text{subject to} \quad & \mathcal{A}(X) = b, \\ & \mathcal{C}(X) + s_c = d, \\ & \mathcal{F}(X) + S_f = G, \quad S_f \succeq 0, \\ & \mathcal{G}(X) + s_g = h, \quad \|s_g\|_\gamma \leq \rho, \end{aligned} \tag{5.2}$$

The FALC framework extended to this more general problem inexactly solves optimization problems of the form:

$$\min_{X, s_c, \|s_g\|_\gamma \leq \rho, S_f \succeq 0} \left\{ \begin{aligned} & \lambda^{(k)} \left(\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s_c\|_\beta + \langle R, X \rangle + \frac{\mu_3}{2} \|X - X_0\|_F^2 \right) \\ & - \lambda^{(k)} (\theta_1^{(k)})^T (\mathcal{A}(X) - b) + \frac{1}{2} \|\mathcal{A}(X) - b\|_2^2 \\ & - \lambda^{(k)} (\theta_2^{(k)})^T (\mathcal{C}(X) + s_c - d) + \frac{1}{2} \|\mathcal{C}(X) + s_c - d\|_2^2 \\ & - \lambda^{(k)} (\theta_3^{(k)})^T (\mathcal{F}(X) + S_f - G) + \frac{1}{2} \|\mathcal{F}(X) + S_f - G\|_2^2 \\ & - \lambda^{(k)} (\theta_4^{(k)})^T (\mathcal{G}(X) + s_g - h) + \frac{1}{2} \|\mathcal{G}(X) + s_g - h\|_2^2 \end{aligned} \right\}$$

Note that we do *not* dualize neither the norm constraint $\|s_g\|_\gamma \leq \rho$ nor the cone constraint $S_f \succeq 0$. FALC solves (5.3) by solving constrained shrinkage problems of the following form.

1. Matrix optimization problem over simple sets:

$$\min_X \left\{ \delta \|\sigma(X)\|_\alpha + \frac{1}{2} \|X - Y\|_F^2 : \|\sigma(X)\|_\alpha \leq \eta \right\}, \quad (5.3)$$

$$\min_{S_f} \left\{ \frac{1}{2} \|S_f - Y\|_F^2 : S_f \succeq 0 \right\}. \quad (5.4)$$

For a given $Y \in \mathbb{R}^{m \times n}$, these problems can be efficiently solved when $\|\sigma(\cdot)\|_\alpha$ is either the nuclear norm, Frobenius norm, or the ℓ_2 -norm, or equivalently, the ℓ_1 , ℓ_2 or ℓ_∞ norm of the singular values of X . Note that subproblem given in (5.4) is only needed when solving the augmented Lagrangian subproblem corresponding to $\mathcal{F}(X) \leq G$ constraints.

2. Vector optimization problem over simple sets:

$$\min_x \left\{ \delta \|x\|_\beta + \frac{1}{2} \|x - y\|_2^2 \right\}, \quad (5.5)$$

$$\min_{s_g} \left\{ \frac{1}{2} \|s_g - y\|_2^2 : \|s_g\|_\gamma \leq \rho \right\}. \quad (5.6)$$

For a given y , these problems can be efficiently solved when β and γ are either ℓ_2 , ℓ_1 or ℓ_∞ vector norms. Note that subproblem given in (5.6) is only needed when solving the augmented Lagrangian subproblem corresponding to $\|\mathcal{G}(X) - h\|_\gamma \leq \rho$ constraints.

The extension (5.1) allows us to model a wider class of problems. Setting $\mathcal{A} = \mathcal{F} = \mathbf{0}$ and $\gamma = 2$ results in a special case that includes matrix completion problems with noisy data. Setting $\beta = \infty$, $\alpha = 2$ dropping the norm constraint $\|\mathcal{G}(X) - h\|_\gamma \leq \rho$, results in a special case that arises in the optimal acquisition basis design for compressive sensing.

The main contribution of this paper is an efficient first-order augmented lagrangian algorithm (FALC) for the composite norm minimization problem (1.1) and by extension (5.1). The FALC recovers the low rank target matrix by solving a sequence of augmented lagrangian subproblems, and each subproblem is solved using Algorithm 3 in [33]. We show that the continuation scheme on penalty parameter λ used in FALC provably converges to the target signal and we are also able to compute a convergence rate. The performance of FALC in our limited numerical experiments has been very promising.

REFERENCES

- [1] N. S. AYBAT AND G. IYENGAR, *A first-order augmented lagrangian method for compressed sensing*, submitted to SIAM Journal on Optimization, (2009).
- [2] ———, *A first-order smoothed penalty method for compressed sensing*, forthcoming in SIAM Journal on Optimization, (2010).
- [3] S. BOYD AND L. VANDENBERGHE, *Convex optimization*, Cambridge University Press, 2004.
- [4] J. CAI, E. CANDÈS, AND Z. SHEN, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 20 (2008), pp. 1956–1982.
- [5] E. CANDÈS AND J. ROMBERG, *Quantitative robust uncertainty principles and optimally sparse decompositions*, Foundations of Computational Mathematics, 6 (2006), pp. 227–254.
- [6] E. CANDÈS, J. ROMBERG, AND T. TAO, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Info. Th., 52 (2006).
- [7] E. CANDÈS AND T. TAO, *Near optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Info. Th., 52 (2006), pp. 5406–5425.
- [8] E. J. CANDÈS, X. LI, Y. MA, AND WRIGHT J., *Robust principle component analysis?*, submitted for publication, (2009).
- [9] A. D’ASPREMONT, F. R. BACH, AND L. EL. GHAOUI, *Optimal solutions for sparse principle component analysis*, Journal of Machine Learning Research, 9 (2008), pp. 1269–1294.
- [10] A. D’ASPREMONT, L. EL. GHAOUI, M. I. JORDAN, AND G. R. G. LANCKRIET, *A direct formulation for sparse pca using semidefinite programming*, SIAM Review, 49 (2007), pp. 434–448.
- [11] I. DAUBECHIES, M. DEFRISE, AND C. DE MOL, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure and Applied Mathematics, 57 (2004), pp. 1413–1457.
- [12] I. DAUBECHIES, M. FORNASIER, AND I. LORIS, *Accelerated projected gradient method for linear inverse problems with sparsity constraints*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 764–792.
- [13] D. DONOHO, *Compressed sensing*, IEEE Trans. Info. Th., 52 (2006), pp. 1289–1306.
- [14] L. EL GHAOUI AND P. GAHINET, *Rank minimization under lmi constraints: A framework for output feedback problems*, in Proceedings of the European Control Conference, 1993.
- [15] M. FAZEL, H. HINDI, AND S. BOYD, *A rank minimization heuristic with application to minimum order system approximation*, in Proceedings of the American Control Conference, 2003, pp. 2156–2162.
- [16] ———, *Rank minimization and applications in system theory*, in American Control Conference, 2004, pp. 3273–3278.
- [17] M. A. FIGUEIREDO, R. NOWAK, AND S. J. WRIGHT, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE Journal of Selected Topics in Signal Processing, 1 (2007), pp. 586–597.
- [18] D. GOLDFARB, S. MA, AND K. SCHEINBERG, *Fast alternating linearization methods for minimizing the sum of two convex functions*. arXiv:0912.4571v2, October 2010.
- [19] E. T. HALE, W. YIN, AND Y. ZHANG, *A fixed-point continuation for ℓ_1 -regularized minimization with applications to compressed sensing*, tech. report, Rice University, 2007.
- [20] ———, *Fixed-point continuation for ℓ_1 -minimization: Methodology and convergence*, SIAM Journal on Optimization, 19 (2008), pp. 1107–1130.
- [21] M. JOURNÉE, Y. NESTEROV, P. RICHTÁRIK, AND SEPULCHRE R., *Generalized power method for sparse principle component analysis*, Journal of Machine Learning Research, 11 (2010), pp. 517–553.
- [22] K. KOH, S. J. KIM, AND S. BOYD, *Solver for ℓ_1 -regularized least squares problems*, tech. report, Stanford University, 2007.
- [23] R.M. LARSEN, *Lanczos bidiagonalization with partial reorthogonalization*, Technical report DAIMI PB-357, Department of Computer Science, Aarhus University, 1998.
- [24] A. S. LEWIS, *The convex analysis of unitarily invariant matrix norms*, Journal of Convex Analysis, 2 (1995), pp. 173–183.
- [25] Z. LIN, M. CHEN, L. WU, AND Y. MA, *The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices*, submitted for publication, (2009).
- [26] Z. LIN, A. GANESH, J. WRIGHT, L. WU, M. CHEN, AND Y. MA, *Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix*, tech. report, UIUC Technical Report UILU-ENG-09-2214, 2009.
- [27] N. LINIAL, E. LONDON, AND Y. RABINOVICH, *The geometry of graphs and some of its algorithmic applications*, Combinatorica, 15 (1995), pp. 215–245.
- [28] Z. LIU AND L. VANDENBERGHE, *Interior-point method for nuclear norm approximation with application to system identification*, Submitted to Mathematical Programming Series B, (2008).
- [29] S. MA, D. GOLDFARB, AND L. CHEN, *Fixed point and bregman iterative methods for matrix rank minimization*, To appear in Mathematical Programming Series A, (2008).
- [30] NETFLIX PRIZE. <http://www.netflixprize.com/>.
- [31] B. RECHT, M. FAZEL, AND P. PARRILO, *Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization*, Submitted to SIAM Review, (2007).
- [32] K.C. TOH AND S. YUN, *An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems*. preprint, 2010.
- [33] P. TSENG, *On accelerated proximal gradient methods for convex-concave optimization*, submitted to SIAM Journal on Optimization, (2008).
- [34] E. VAN DEN BERG AND M. P. FRIEDLANDER, *Probing the pareto frontier for basis pursuit solutions*, SIAM Journal on Scientific Computing, 31 (2008), pp. 890–912.
- [35] Z. WEN, W. YIN, D. GOLDFARB, AND Y. ZHANG, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation*, tech. report, Columbia University, 2009.

- [36] W. YIN, S. OSHER, D. GOLDFARB, AND J. DARON, *Bregman iterative algorithms for ℓ_1 minimization with applications to compressed sensing*, SIAM Journal on Imaging Sciences, 1 (2008), pp. 143–168.
- [37] Z. ZHOU, X. LI, J. WRIGHT, E. CANDÈS, AND Y. MA, *Stable principle component pursuit*, in Proceedings of International Symposium on Information Theory, 2010.

Appendix A. Auxiliary results.

THEOREM A.1. *Let $f : \mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ denote a convex function with a Lipschitz continuous gradient ∇f with a Lipschitz constant L with respect to the norm $\|\cdot\|$ on $\mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q$ defined as follows:*

$$\|(X, s, y)\| = \sqrt{\|X\|_F^2 + \|s\|_2^2 + \|y\|_2^2}. \quad (\text{A.1})$$

Let $(X_, s_*, y_*) \in \operatorname{argmin}\{\lambda(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + f(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$. Suppose $(\bar{X}, \bar{s}, \bar{y}) \in \mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q$ such that $\|\bar{y}\|_\gamma \leq \rho$ satisfies*

$$\lambda(\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\bar{s}\|_\beta) + f(\bar{X}, \bar{s}, \bar{y}) \leq \lambda(\mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|s_*\|_\beta) + f(X_*, s_*, y_*) + \epsilon$$

for some $\epsilon > 0$. Then

$$\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F \leq (\sqrt{2L\epsilon} + I(\alpha^*)\lambda\mu_1), \quad \|\nabla_s f(\bar{X}, \bar{s}, \bar{y})\|_2 \leq (\sqrt{2L\epsilon} + J(\beta^*)\lambda\mu_2)$$

where $I(\cdot)$ and $J(\cdot)$ are defined in (2.4), α^ and β^* denote the Hölder conjugate of α and β , respectively.*

Proof. Since ∇f is Lipschitz continuous with constant L , the triangular inequality for $\|\sigma(\cdot)\|_\alpha$ and $\|\cdot\|_\beta$ implies that for any $Y \in \mathbb{R}^{m \times n}$, $q \in \mathbb{R}^p$ and $z \in \mathbb{R}^q$

$$\begin{aligned} \lambda(\mu_1\|\sigma(Y)\|_\alpha + \mu_2\|q\|_\beta) + f(Y, q, z) &\leq \lambda(\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\bar{s}\|_\beta) + f(\bar{X}, \bar{s}, \bar{y}) + \lambda(\mu_1\|\sigma(Y - \bar{X})\|_\alpha + \mu_2\|q - \bar{s}\|_\beta) \\ &\quad + \langle \nabla_X f(\bar{X}, \bar{s}, \bar{y}), (Y - \bar{X}) \rangle + \nabla_s f(\bar{X}, \bar{s}, \bar{y})^T (q - \bar{s}) + \nabla_y f(\bar{X}, \bar{s}, \bar{y})^T (z - \bar{y}) \\ &\quad + \frac{L}{2}\|Y - \bar{X}\|_F^2 + \frac{L}{2}\|q - \bar{s}\|_2^2 + \frac{L}{2}\|z - \bar{y}\|_2^2, \end{aligned}$$

where $\langle X, Y \rangle = \operatorname{Tr}(X^T Y) \in \mathbb{R}$ denotes the usual Euclidean inner product of $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{m \times n}$. Since Y , q and z are arbitrary, it follows that

$$\begin{aligned} \lambda(\mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|s_*\|_\beta) + f(X_*, s_*, y_*) &\leq \lambda(\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\bar{s}\|_\beta) + f(\bar{X}, \bar{s}, \bar{y}) \\ &\quad + \min_{Y \in \mathbb{R}^{m \times n}} \left\{ \langle \nabla_X f(\bar{X}, \bar{s}, \bar{y}), Y - \bar{X} \rangle + \frac{L}{2}\|Y - \bar{X}\|_F^2 + \lambda\mu_1\|\sigma(Y - \bar{X})\|_\alpha \right\} \\ &\quad + \min_{q \in \mathbb{R}^p} \left\{ \nabla_s f(\bar{X}, \bar{s}, \bar{y})^T (q - \bar{s}) + \frac{L}{2}\|q - \bar{s}\|_2^2 + \lambda\mu_2\|q - \bar{s}\|_\beta \right\} \\ &\quad + \min_{z: \|z\|_\gamma \leq \rho} \left\{ \nabla_y f(\bar{X}, \bar{s}, \bar{y})^T (z - \bar{y}) + \frac{L}{2}\|z - \bar{y}\|_2^2 \right\}. \end{aligned} \quad (\text{A.2})$$

The first minimization problem on the right hand side of (A.2) can be simplified as follows:

$$\begin{aligned} &\min_{Y \in \mathbb{R}^{m \times n}} \left\{ \langle \nabla_X f(\bar{X}, \bar{s}, \bar{y}), Y - \bar{X} \rangle + \frac{L}{2}\|Y - \bar{X}\|_F^2 + \lambda\mu_1\|\sigma(Y - \bar{X})\|_\alpha \right\} \\ &= \max_{W: \|\sigma(W)\|_{\alpha^*} \leq \lambda\mu_1} \min_{Y \in \mathbb{R}^{m \times n}} \left\{ \frac{L}{2}\|Y - \bar{X}\|_F^2 + \langle \nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W, Y - \bar{X} \rangle \right\}, \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} &= \max_{W: \|\sigma(W)\|_{\alpha^*} \leq \lambda\mu_1} \left\{ \frac{L}{2}\|Y^*(W) - \bar{X}\|_F^2 + \langle \nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W, Y^*(W) - \bar{X} \rangle \right\}, \\ &= - \min_{W: \|\sigma(W)\|_{\alpha^*} \leq \lambda\mu_1} \frac{\|\nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W\|_F^2}{2L}, \end{aligned} \quad (\text{A.4})$$

$Y^*(W) = \bar{X} - \frac{\nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W}{L}$ is the minimizer of the inner minimization problem in (A.3).

The second minimization problem on the right hand side of (A.2) can be simplified as follows:

$$\begin{aligned} & \min_{q \in \mathbb{R}^p} \left\{ \nabla_s f(\bar{X}, \bar{s}, \bar{y})^T (q - \bar{s}) + \frac{L}{2} \|q - \bar{s}\|_2^2 + \lambda \mu_2 \|q - \bar{s}\|_\beta \right\} \\ &= \max_{u: \|u\|_{\beta^*} \leq \lambda \mu_2} \min_{q \in \mathbb{R}^p} \left\{ \frac{L}{2} \|q - \bar{s}\|_2^2 + (\nabla_s f(\bar{X}, \bar{s}, \bar{y}) + u)^T (q - \bar{s}) \right\}, \end{aligned} \quad (\text{A.5})$$

$$\begin{aligned} &= \max_{u: \|u\|_{\beta^*} \leq \lambda \mu_2} \left\{ \frac{L}{2} \|q^*(u) - \bar{s}\|_2^2 + (\nabla_s f(\bar{X}, \bar{s}, \bar{y}) + u)^T (q^*(u) - \bar{s}) \right\}, \\ &= - \min_{u: \|u\|_{\beta^*} \leq \lambda \mu_2} \frac{\|\nabla_s f(\bar{X}, \bar{s}, \bar{y}) + u\|_2^2}{2L}, \end{aligned} \quad (\text{A.6})$$

$q^*(u) = \bar{s} - \frac{\nabla_s f(\bar{X}, \bar{s}, \bar{y}) + u}{L}$ is the minimizer of the inner minimization problem in (A.5).

Since $\|\bar{y}\|_\gamma \leq \rho$, then the following is true for the third minimization problem on the right hand side of (A.2).

$$\min_{z: \|z\|_\gamma \leq \rho} \left\{ \nabla_y f(\bar{X}, \bar{s}, \bar{y})^T (z - \bar{y}) + \frac{L}{2} \|z - \bar{y}\|_2^2 \right\} \leq 0. \quad (\text{A.7})$$

Thus, (A.2), (A.4), (A.6) and (A.7) together imply that

$$\begin{aligned} \lambda(\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|s_*\|_\beta) + f(X_*, s_*, y_*) &\leq \lambda(\mu_1 \|\sigma(\bar{X})\|_\alpha + \mu_2 \|\bar{s}\|_\beta) + f(\bar{X}, \bar{s}, \bar{y}) \\ &\quad - \min_{W: \|\sigma(W)\|_{\alpha^*} \leq \lambda \mu_1} \frac{\|\nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W\|_F^2}{2L} \\ &\quad - \min_{u: \|u\|_{\beta^*} \leq \lambda \mu_2} \frac{\|\nabla_s f(\bar{X}, \bar{s}, \bar{y}) + u\|_2^2}{2L}. \end{aligned}$$

Since $\left(\lambda(\mu_1 \|\sigma(\bar{X})\|_\alpha + \mu_2 \|\bar{s}\|_\beta) + f(\bar{X}, \bar{s}, \bar{y}) \right) - \left(\lambda(\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|s_*\|_\beta) + f(X_*, s_*, y_*) \right) \leq \epsilon$, we have that

$$\min_{W: \|\sigma(W)\|_{\alpha^*} \leq \lambda \mu_1} \|\nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W\|_F^2 + \min_{u: \|u\|_{\beta^*} \leq \lambda \mu_2} \|\nabla_s f(\bar{X}, \bar{s}, \bar{y}) + u\|_2^2 \leq 2L\epsilon. \quad (\text{A.8})$$

From the definition of $I(\cdot)$ in (2.4), it follows that $\|W\|_F \leq I(\alpha^*) \|\sigma(W)\|_{\alpha^*}$. Thus, (A.8) implies that

$$\min_{W: \|W\|_F \leq I(\alpha^*) \lambda \mu_1} \|\nabla_X f(\bar{X}, \bar{s}, \bar{y}) + W\|_F^2 \leq 2L\epsilon. \quad (\text{A.9})$$

Suppose $\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F > I(\alpha^*) \lambda \mu_1$. Then the optimal solution of the optimization problem in (A.9) is

$$W^* = -I(\alpha^*) \lambda \mu_1 \cdot \frac{\nabla_X f(\bar{X}, \bar{s}, \bar{y})}{\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F}.$$

Then (A.8) implies that $(\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F - I(\alpha^*) \lambda \mu_1)^2 \leq 2L\epsilon$, i.e. $\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F \leq \sqrt{2L\epsilon} + I(\alpha^*) \lambda \mu_1$. This is trivially true when $\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F \leq I(\alpha^*) \lambda \mu_1$. Therefore, we can conclude that always

$$\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F \leq \sqrt{2L\epsilon} + I(\alpha^*) \lambda \mu_1.$$

A similar analysis establishes that $\|\nabla_s f(\bar{X}, \bar{s}, \bar{y})\|_2 \leq \sqrt{2L\epsilon} + J(\beta^*) \lambda \mu_2$. \square

COROLLARY A.2. Let $\alpha, \beta \in \{1, 2, \infty\}$ and

$$P(X, s, y) = \lambda(\mu_1 \|\sigma(X)\|_\alpha + \mu_2 \|s\|_\beta) + f(X, s, y), \quad f(X, s, y) = \frac{1}{2} \|\mathcal{A}(X) + y - b - \lambda \theta_1\|_2^2 + \frac{1}{2} \|\mathcal{C}(X) + s - d - \lambda \theta_2\|_2^2,$$

where $\mathcal{A}: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^q$ and $\mathcal{C}: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^p$ denote linear $\mathcal{A}(X) = A \mathbf{vec}(X)$, $\mathcal{C}(X) = C \mathbf{vec}(X)$ for all $X \in \mathbb{R}^{m \times n}$, where $A \in \mathbb{R}^{q \times mn}$ and $C \in \mathbb{R}^{p \times mn}$ are the matrix representation of the linear maps $\mathcal{A}(\cdot)$ and $\mathcal{C}(\cdot)$, respectively; and $\mathbf{vec}(X)$ denotes the vector obtained by stacking the columns of X in order.

Suppose $(\bar{X}, \bar{s}, \bar{y})$ is ϵ -optimal for the problem $\min\{P(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$, i.e.

$$0 \leq P(\bar{X}, \bar{s}, \bar{y}) - \min_{X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho} P(X, s, y) \leq \epsilon,$$

and the matrix A has full row rank. Then

$$\begin{aligned} \|\mathcal{A}(\bar{X}) + \bar{y} - b - \lambda\theta_1\|_2 &\leq \frac{I(\alpha^*)\mu_1 + \sigma_{\max}(C)J(\beta^*)\mu_2}{\sigma_{\min}(A)}\lambda + \frac{\sigma_{\max}(M)(1 + \sigma_{\max}(C))}{\sigma_{\min}(A)}\sqrt{2}\epsilon, \\ \|\mathcal{C}(\bar{X}) + \bar{s} - d - \lambda\theta_2\|_2 &\leq J(\beta^*)\mu_2\lambda + \sigma_{\max}(M)\sqrt{2}\epsilon, \end{aligned}$$

where $\sigma_{\max}(X)$ and $\sigma_{\min}(X)$ denote respectively the maximum and minimum singular values of a matrix X ;

$M = \begin{pmatrix} I & 0 & C \\ 0 & I & A \end{pmatrix}$, and $I(\alpha^*)$, $J(\beta^*)$ are defined in (2.4).

Proof. Let $f(X, s, y) = \frac{1}{2}\|\mathcal{A}(X) + y - b - \lambda\theta_1\|_2^2 + \frac{1}{2}\|\mathcal{C}(X) + s - d - \lambda\theta_2\|_2^2$ and $\|\cdot\|$ be the norm on $\mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q$ defined in (A.1), then for any $X_1, X_2 \in \mathbb{R}^{m \times n}$, $s_1, s_2 \in \mathbb{R}^p$ and $y_1, y_2 \in \mathbb{R}^q$, we have

$$\begin{aligned} &\|\nabla f(X_1, s_1, y_1) - \nabla f(X_2, s_2, y_2)\|^2 \\ &= \|\nabla_X f(X_1, s_1, y_1) - \nabla_X f(X_2, s_2, y_2), \nabla_s f(X_1, s_1, y_1) - \nabla_s f(X_2, s_2, y_2), \nabla_y f(X_1, s_1, y_1) - \nabla_y f(X_2, s_2, y_2)\|^2, \\ &= \|\nabla_X f(X_1, s_1, y_1) - \nabla_X f(X_2, s_2, y_2)\|_F^2 + \|\nabla_s f(X_1, s_1, y_1) - \nabla_s f(X_2, s_2, y_2)\|_2^2 + \|\nabla_y f(X_1, s_1, y_1) - \nabla_y f(X_2, s_2, y_2)\|_2^2, \\ &= \|\mathcal{A}^*(\mathcal{A}(X_1 - X_2) + y_1 - y_2) + \mathcal{C}^*(\mathcal{C}(X_1 - X_2) + s_1 - s_2)\|_F^2 \\ &\quad + \|\mathcal{C}(X_1 - X_2) + s_1 - s_2\|_2^2 + \|\mathcal{A}(X_1 - X_2) + y_1 - y_2\|_2^2, \\ &= \|A^T(A \text{vec}(X_1 - X_2) + y_1 - y_2) + C^T(C \text{vec}(X_1 - X_2) + s_1 - s_2)\|_2^2 \\ &\quad + \|C \text{vec}(X_1 - X_2) + s_1 - s_2\|_2^2 + \|A \text{vec}(X_1 - X_2) + y_1 - y_2\|_2^2, \\ &= \|M^T M \begin{pmatrix} s_1 - s_2 \\ y_1 - y_2 \\ \text{vec}(X_1 - X_2) \end{pmatrix}\|_2^2. \end{aligned}$$

Hence,

$$\begin{aligned} \|\nabla f(X_1, s_1, y_1) - \nabla f(X_2, s_2, y_2)\| &\leq \sigma_{\max}^2(M) \left\| \begin{pmatrix} s_1 - s_2 \\ y_1 - y_2 \\ \text{vec}(X_1 - X_2) \end{pmatrix} \right\|_2, \\ &= \sigma_{\max}^2(M) \sqrt{\|X_1 - X_2\|_F^2 + \|s_1 - s_2\|_2^2 + \|y_1 - y_2\|_2^2}, \\ &= \sigma_{\max}^2(M) \|(X_1, s_1, y_1) - (X_2, s_2, y_2)\|, \end{aligned}$$

where $\sigma_{\max}(M)$ is the maximum singular-value of M . Thus, $f : \mathbb{R}^{m \times n} \times \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ is a convex function and ∇f is Lipschitz continuous with respect to $\|\cdot\|$ with Lipschitz constant $L = \sigma_{\max}^2(M)$.

Since $(\bar{X}, \bar{s}, \bar{y})$ is an ϵ -optimal solution to the problem $\min\{P(X, s, y) : X \in \mathbb{R}^{m \times n}, s \in \mathbb{R}^p, \|y\|_\gamma \leq \rho\}$, Theorem A.1 guarantees that

$$\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F = \|\mathcal{A}^*(\mathcal{A}(\bar{X}) + \bar{y} - b - \lambda\theta_1) + \mathcal{C}^*(\mathcal{C}(\bar{X}) + \bar{s} - d - \lambda\theta_2)\|_F \leq \sqrt{2}\epsilon \sigma_{\max}(M) + I(\alpha^*)\lambda\mu_1, \quad (\text{A.10})$$

and

$$\|\nabla_s f(\bar{X}, \bar{s}, \bar{y})\|_2 = \|\mathcal{C}(\bar{X}) + \bar{s} - d - \lambda\theta_2\|_2 \leq \sqrt{2}\epsilon \sigma_{\max}(M) + J(\beta^*)\lambda\mu_2. \quad (\text{A.11})$$

The bound (A.10) and the triangular inequality for Frobenius norm implies that

$$\|\mathcal{A}^*(\mathcal{A}(\bar{X}) + \bar{y} - b - \lambda\theta_1)\|_F \leq \|\mathcal{C}^*(\mathcal{C}(\bar{X}) + \bar{s} - d - \lambda\theta_2)\|_F + \sqrt{2}\epsilon \sigma_{\max}(M) + I(\alpha^*)\lambda\mu_1. \quad (\text{A.12})$$

Since $\|\mathcal{C}^*(\mathcal{C}(\bar{X}) + \bar{s} - d - \lambda\theta_2)\|_F \leq \sigma_{\max}(C) \|\mathcal{C}(\bar{X}) + \bar{s} - d - \lambda\theta_2\|_2$, (A.12) implies that

$$\|\mathcal{A}^*(\mathcal{A}(\bar{X}) - \bar{y} - b - \lambda\theta_1)\|_F \leq \sigma_{\max}(C) (\sqrt{2\epsilon}\sigma_{\max}(M) + J(\beta^*)\lambda\mu_2) + \sqrt{2\epsilon}\sigma_{\max}(M) + I(\alpha^*)\lambda\mu_1.$$

Consequently,

$$\begin{aligned} \|\mathcal{A}(\bar{X}) + \bar{y} - b - \lambda\theta_1\|_2 &\leq \frac{1}{\sigma_{\min}(A)} \cdot \|\mathcal{A}^*(\mathcal{A}(\bar{X}) + \bar{y} - b - \lambda\theta_1)\|_F, \\ &\leq \frac{1}{\sigma_{\min}(A)} \cdot \left(\sigma_{\max}(C) (\sqrt{2\epsilon}\sigma_{\max}(M) + J(\beta^*)\lambda\mu_2) + \sqrt{2\epsilon}\sigma_{\max}(M) + I(\alpha^*)\lambda\mu_1 \right). \end{aligned}$$

□

LEMMA A.3. Let $(\mathcal{M}, \|\cdot\|)$ be a normed vector space, $f : \mathcal{M} \rightarrow \mathbb{R}$ be a strictly convex function and $\chi \subset \mathcal{M}$ be a closed, convex set with a non-empty interior. Let $\bar{x} = \operatorname{argmin}_{x \in \chi} f(x)$ and $x^* = \operatorname{argmin}_{x \in \mathcal{M}} f(x)$. If $x^* \notin \chi$, then $\bar{x} \in \mathbf{bd} \chi$, where $\mathbf{bd} \chi$ denotes the boundary of χ .

Proof. We will establish the result by contradiction. Assume \bar{x} is in the interior of χ , i.e. $\bar{x} \in \mathbf{int}(\chi)$. Then $\exists \epsilon > 0$ such that $B(\bar{x}, \epsilon) = \{x \in \mathcal{M} : \|x - \bar{x}\| < \epsilon\} \subset \chi$. Since f is strictly convex and $x^* \neq \bar{x}$, $f(x^*) < f(\bar{x})$. Choose $0 < \lambda < \frac{\epsilon}{\|\bar{x} - x^*\|} < 1$ so that $\lambda x^* + (1 - \lambda)\bar{x} \in B(\bar{x}, \epsilon) \subset \chi$. Since f is strictly convex,

$$f(\lambda x^* + (1 - \lambda)\bar{x}) < \lambda f(x^*) + (1 - \lambda)f(\bar{x}) < f(\bar{x}). \quad (\text{A.13})$$

However, $\lambda x^* + (1 - \lambda)\bar{x} \in B(\bar{x}, \epsilon) \subset \chi$ and $f(\lambda x^* + (1 - \lambda)\bar{x}) < f(\bar{x})$ contradicts the fact that $f(\bar{x}) < f(x)$ for all $x \in \chi$. Therefore, $\bar{x} \notin \mathbf{int}(\chi)$. Since $\bar{x} \in \chi$, it follows that $\bar{x} \in \mathbf{bd} \chi$. □

Next, we collect together complexity results for optimization problems of the form $\min_{X \in \mathbb{R}^{m \times n}} \{\delta \|\sigma(X)\|_\alpha + \frac{1}{2} \|X - Y\|_F^2 : \|\sigma(X)\|_\alpha \leq \eta\}$ and $\min_{s \in \mathbb{R}^p} \{\delta \|s\|_\beta + \frac{1}{2} \|s - q\|_2^2 : \|s\|_\beta \leq \eta\}$ that need to be solved in each FALC update step.

LEMMA A.4. Let $X^* = \operatorname{argmin} \{\delta \|\sigma(X)\|_\alpha + \frac{1}{2} \|X - Y\|_F^2 : \|\sigma(X)\|_\alpha \leq \eta, X \in \mathbb{R}^{m \times n}\}$ of the constrained matrix shrinkage problem. Then

$$X^* = U \operatorname{diag}(s^*) V^T,$$

where $U \operatorname{diag}(\sigma) V^T$, $\sigma \in \mathbb{R}_+^r$, denotes the SVD of Y and s^* denotes the optimal solution of the constrained vector shrinkage problem

$$\min \left\{ \delta \|s\|_\alpha + \frac{1}{2} \|s - \sigma\|_2^2 : \|s\|_\alpha \leq \eta, s \in \mathbb{R}^r \right\}.$$

Since the worst case complexity of computing the SVD of Y is $\mathcal{O}(\min\{n^2 m, m^2 n\})$ the complexity of the computing X^* is $\mathcal{O}(\min\{n^2 m, m^2 n\} + T_v(r, \alpha))$, where $T_v(r, \alpha)$ denotes the complexity of computing the solution of an r -dimensional constrained vector shrinkage problem with norm $\|\cdot\|_\alpha$. The function

$$T_v(p, \alpha) = \begin{cases} \mathcal{O}(p \ln(p)) & \alpha = 1, \infty, \\ \mathcal{O}(p), & \alpha = 2, \end{cases} \quad (\text{A.14})$$

Proof. The standard results in non-linear convex optimization over matrices implies that X^* is of the form $X^* = U \operatorname{diag}(s^*) V^T$ (see Corollary 2.5 in [24]).

Now, consider the vector constrained shrinkage problem $\min_{x \in \mathbb{R}^p} \{\delta \|x\|_\beta + \frac{1}{2} \|x - y\|_2^2 : \|x\|_\beta \leq \eta\}$.

(i) $\beta = 1$: See Lemma A.4 in [1].

(ii) $\beta = 2$: First considered the unconstrained case, i.e. $\eta = \infty$. Since ℓ_2 -norm is self dual, $\delta \|x\|_2 = \max\{u^T x : \|u\|_2 \leq 1\}$. Thus,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \left\{ \delta \|x\|_2 + \frac{1}{2} \|x - y\|_2^2 \right\} &= \min_{x \in \mathbb{R}^n} \max_{u : \|u\|_2 \leq \delta} \left\{ u^T x + \frac{1}{2} \|x - y\|_2^2 \right\}, \\ &= \max_{u : \|u\|_2 \leq \delta} \min_{x \in \mathbb{R}^n} \left\{ u^T x + \frac{1}{2} \|x - y\|_2^2 \right\}, \\ &= \max_{u : \|u\|_2 \leq \delta} \left\{ u^T (u - y) + \frac{1}{2} \|u\|_2^2 \right\}, \\ &= \frac{1}{2} \|y\|_2^2 - \min_{u : \|u\|_2 \leq \delta} \frac{1}{2} \|u - y\|_2^2, \end{aligned} \quad (\text{A.15})$$

where (A.15) follows from the fact that $x^*(u) := \operatorname{argmin}_x u^T x + \frac{1}{2} \|x - y\|_2^2 = y - u$. Define

$$u^* := \operatorname{argmin}_{u: \|u\|_2 \leq \delta} \frac{1}{2} \|u - y\|_2^2 = y \min \left\{ \frac{\delta}{\|y\|_2}, \mathbf{1} \right\}.$$

Then the unconstrained optimal solution $\bar{x} = x^*(u^*) = y \max \left\{ 1 - \frac{\delta}{\|y\|_2}, 0 \right\}$ and the complexity of computing \bar{x} is $\mathcal{O}(p)$.

Next, consider the constrained optimization problem. The constrained optimum $x^* = \bar{x}$, whenever \bar{x} is feasible, i.e. $\|\bar{x}\|_2 \leq \beta$. Since $f(x) := \delta \|x\|_2 + \frac{1}{2} \|x - y\|_2^2$ is strongly convex, Lemma A.3 implies that $\|x^*\|_2 = \eta$ whenever $\|\bar{x}\|_2 > \eta$. Thus,

$$\min \left\{ \delta \|x\|_2 + \frac{1}{2} \|x - y\|_2^2 : \|x\|_2 \leq \eta \right\} = \delta \eta + \min \left\{ \frac{1}{2} \|x - y\|_2^2 : \|x\|^2 = \eta^2 \right\}.$$

The unique KKT point for the optimization problem $\min \left\{ \frac{1}{2} \|x - y\|_2^2 : \|x\|^2 = \eta^2 \right\}$, is given by $x^* = \frac{\eta}{\|y\|} y$ and KKT multiplier for the constraint $\|x\|^2 = \eta^2$ is $\vartheta = \frac{\|y\|_2}{\beta} - 1$. It is easy to check that $\vartheta > 0$ whenever $\|\bar{x}\|_2 > \beta$. Thus, x^* is optimal for the convex optimization problem $\min \left\{ \min \left\{ \frac{1}{2} \|x - y\|_2^2 : \|x\|^2 \leq \eta^2 \right\} \right\}$; consequently, optimal for equality constrained optimization problem $\min \left\{ \frac{1}{2} \|x - y\|_2^2 : \|x\| = \eta \right\}$. Hence, the complexity of computing x^* is $\mathcal{O}(p)$.

- (iii) $\beta = \infty$: First consider the unconstrained problem. Since ℓ_1 -norm is the dual norm of the ℓ_∞ -norm, we have that

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \left\{ \delta \|x\|_\infty + \frac{1}{2} \|x - y\|_2^2 \right\} &= \min_{x \in \mathbb{R}^n} \max_{u: \|u\|_1 \leq \delta} \left\{ u^T x + \frac{1}{2} \|x - y\|_2^2 \right\}, \\ &= \max_{u: \|u\|_1 \leq \delta} \min_{x \in \mathbb{R}^n} \left\{ u^T x + \frac{1}{2} \|x - y\|_2^2 \right\}, \\ &= \max_{u: \|u\|_1 \leq \delta} \left\{ u^T (y - u) + \frac{1}{2} \|u\|_2^2 \right\}, \\ &= \frac{1}{2} \|y\|_2^2 - \min_{u: \|u\|_1 \leq \delta} \frac{1}{2} \|u - y\|_2^2, \end{aligned} \tag{A.16}$$

where (A.16) follows from the fact that $\operatorname{argmin}_x u^T x + \frac{1}{2} \|x - y\|_2^2 = y - u$. The result in (i) now implies that complexity of computing $u^* = \min_{u: \|u\|_1 \leq \delta} \frac{1}{2} \|u - y\|_2^2$ with $\mathcal{O}(n \log(n))$. Thus, the unconstrained optimal solution $\bar{x} = x^*(u^*) = y - u^*$ can be computed in $\mathcal{O}(p \log(p))$ operations.

Next, consider the constrained optimization problem. The constrained optimum, $x^* = \bar{x}$ whenever \bar{x} is feasible, i.e. $\|\bar{x}\|_\infty \leq \eta$. Since $f(x) = \lambda \|x\|_\infty + \frac{1}{2} \|x - y\|_2^2$ is strictly convex, Lemma A.3 implies that $\|x^*\|_\infty = \eta$, whenever $\|\bar{x}\|_\infty > \eta$. Therefore,

$$\min \left\{ \lambda \|x\|_\infty + \frac{1}{2} \|x - y\|_2^2 : \|x\|_\infty \leq \eta \right\} = \lambda \eta + \min \left\{ \frac{1}{2} \|x - y\|_2^2 : \|x\|_\infty = \eta \right\}$$

It is easy to check $\operatorname{sign}(x_i^*) = \operatorname{sign}(y_i)$ for all $i = 1, \dots, n$. Thus,

$$\min \left\{ \frac{1}{2} \|x - y\|_2^2 : \|x\|_\infty = \eta \right\} = \min \left\{ \frac{1}{2} \|x - |y|\|_2^2 : 0 \leq x_i \leq \eta \right\}.$$

Note that we are guaranteed that $\max_i \{x_i\} = \eta$ because $\|\bar{x}\|_\infty > \eta$. The optimal solution of the 1-dimensional problem $\operatorname{argmin} \left\{ \frac{1}{2} (x - |y|)^2 : 0 \leq x \leq \eta \right\} = \min\{|y|, \eta\}$. Thus, it follows that $x^* = \operatorname{sign}(y) \odot \min\{|y|, \eta \mathbf{1}\}$, \odot denotes componentwise multiplication and $\mathbf{1}$ is a vector of ones, and the complexity of computing x^* is $\mathcal{O}(p \ln(p))$.

□